Summary of changes

#	Change Requested by the supervisor	**How** you addressed the supervisor request
1	Reduce the plagiarism	Reduced the plagiarism up to 19%

Forecast Electricity Sales in Industrial Sector in Sri Lanka Using Predictive Analytics

A. U. B. Chandrasena

2021

Forecast Electricity Sales in Industrial Sector in Sri Lanka Using Predictive Analytics

A dissertation submitted for the Degree of Master of Business Analytics

A. U. B. Chandrasena

University of Colombo School of Computing

2021

DECLARATION

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: A.U.B. Chandrasena

Registration Number: 2018/BA/007

Index Number: 18880072

Signature: Date: 9/13/2021

This is to certify that this thesis is based on the work of

~~Mr~~./Ms. A.U.B. Chandrasena

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: R.A.B. Abeygunawardana

Signature: Date:

I wish to devote this thesis to my parents and sister.

ACKNOWLEDGEMENTS

Above all else, my true much gratitude goes to the University of Colombo School of Computing for choosing me to follow this degree of Master of Business Analytics.

I am incredibly grateful to my supervisor, Dr. R.A.B. Abeygunawardana, for the valuable guidance and support given to me to make this research a success. His vision, sincerity, and motivation have deeply inspired me.

As my course coordinators, my special thanks go to Dr. T. Halloluwa and Dr. M. Wickramasinghe for the support given to me throughout the course.

In collecting data, the support, I received from Mr. Shantha Fonseka of the Ceylon Electricity Board, officers from the Central Bank of Sri Lanka is highly appreciated.

I am very much grateful to My family, and Ms. Hansani Chandrasena and Mr. Oshada Rodrigo, for the kind co-operation and the support given to me in numerous ways in doing this research and the degree.

ABSTRACT

Electricity has turned in to a significant type of end-use energy in the today's advanced society. The impact of electricity is enormous and has been perceived as a fundamental day today need of human.

Forecasting of Electricity sales is significant and critical for a utility to decide on the correct selection relating to future power generation stations and organizational strengthening.

During the last decade, several techniques are being used to forecast electricity sales. This study attempts to review the time series, Autoregressive Moving-Average (ARMA) and Linear Regression methods and choose the most suitable forecasting method for long-term electricity sales forecast using annual data from 1969 to 2018.

The two models were created and fine-tuned using recorded industrial electricity sales of Sri Lanka, and ARIMA (0,2,1) was observed as the best fit model for forecasting annual industrial electricity sales of the Sri Lanka power system with the lowest RMSE of 176.553 GWH, MAPE of 15.42% and R^2 88%.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Background of the study

1.1.1 History of Electricity in Sri Lanka

In 1890, the very first time an electrical bulb was provided electric power was when Sri Lanka was a colony of Britain known as Ceylon. In 1895, a company Messrs Boustead Bros, in Colombo provide electricity publicly for the very first time in Sri Lanka. In 1918, D.J. wimalasurendra an engineer, who later turned into colonizer of power in Sri Lanka, discovered the potential of hydro-power in Sri Lanka near the central hills. In 1926, with the establishment of a separate electricity department, it became possible to meet the expanding demand for electricity. In 1969**,** Ceylon Electricity Board (CEB) was entrenched underneath Parliament Act-No: 17. Since then, CEB is involved in generating, transmitting, and distributing electricity to every consumer category and generate income. It has been also engaged in securing resources, as well as human resources, after the affirmed methodology. CEB must utilize the assets by applying commonsense and dependable administrative techniques (*CEB*).

The Sri Lankan power system has an all-out introduced limit of around 4,048MW by the end of 2018 with a total executable range of 3,464 MW. The recorded outrageous requirements in 2018 were 2616 MW, and the whole net cohort was 15,305GWh (*the Plan for Long-Term Generation Expansion 2018 to 2037,* 2018).

1.1.2 Electricity and Economy

Sri Lanka, a country that is surrounded by waters of the Indian Ocean from all sides, is transforming into an upper-middle-income frugality. The GDP of the country is the financial or hawk cost of the respective abundance of perfect employment and goods designed within a country in a given period. As a wide percentage of overall speaking domestic production, its ranges as a considerable catalog of a specified country's economic health. The economic growth of Sri Lanka on annual basis was 5.5%for the period from 2001 to 2017. The growth of 2017 and 2018 was lower by 3.3% in comparison to the previous year (*GDP growth (annual %) - Sri Lanka | data*). The per capita GDP of Sri Lanka for the year 2017 was $4,080.57 in 2018(*GDP per capita (current US$) - Sri Lanka* | data). The highest subsidy to GDP (56.7%) was achieved by the helping zone; the commercial zone added 26.9%, and farming parts added just 6.8%. The key productive signs for the year 2018 are shown in Table 1.1.

Table 1. 1 Key Economic Indicators, 2018

Indicator	Value
Land area	65610 Km^2
Population	21.67 million
Population Density	346 Persons/km^2
GDP per capita	$4080.57

Source: Economic and Social Statistics of Sri Lanka 2018.

Electricity sales growth in history has often disclosed a straight relationship in comparison to the rate of growth in savings of a country. Figure 1.1 depicts the sales of electricity and the GDP from the year 31 years from 1994.

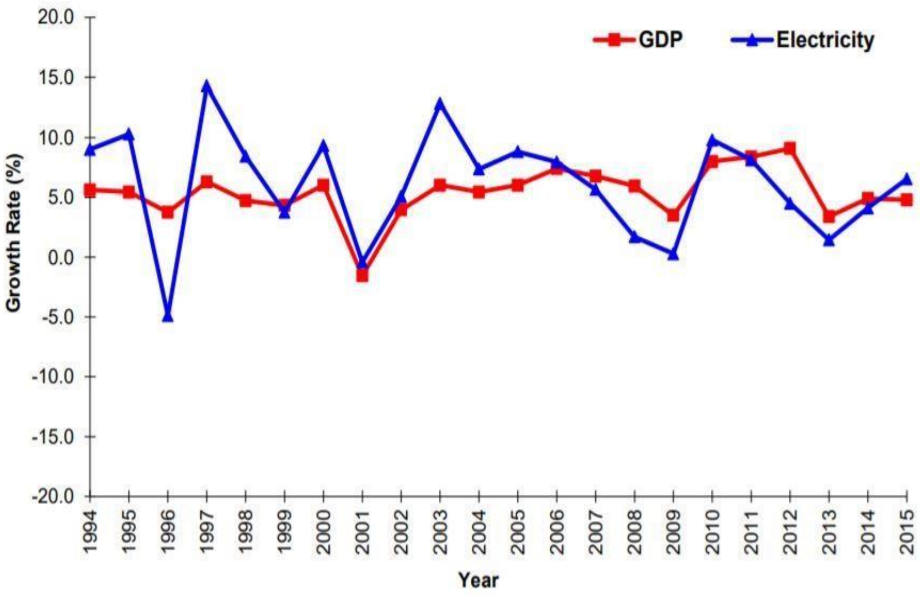


Figure 1. 1 Gross Domestic Product and Electricity Sales

Source: CEB web site

By the above figure, the sales of electricity show a direct relation to the rate at which the economy of the country is growing. In 1996 the growth of electricity shows the lowest from the entire period and in 1997 the very next year shows the highest electricity growth.

1.1.3 Energy Supply and Electricity

According to the 2017 report of the Sri Lanka Energy Balance by Sri Lanka maintainable Energy administration, petroleum, biomass, hydro and coal are significant and essential sources of energy supply, having per unit of population utilization of about 0.5 Tons of Oil Equivalent for the energy of Sri Lanka. Biomass also known as fuelwood, is a primary fuel type that is non-profitable, gives around 40% of the nation's absolute energy requirement. Petroleum ends up being the significant wellspring of commercial energy, covering around 40% of the energy requirement (Sri Lanka Sustainable Energy Authority, 2017).

Even though petroleum and electricity goods are one the fundamental profit-oriented fuels, and growing measure of biomass fuel is now also being used commercially. Hydro-power, which fulfills almost 9% of the total energy needs, is the basic first producer of vital energy that is commercially available in the country. The capacity of the hydro resource is estimated to be about 2,000MW, and many parts of it are being generated. Also, abuse of hydro sources is getting gradually bothersome due to social and, additionally, natural reactions connected to huge scope improvement. Apart from these, there is a remarkable capability for wind and sun energy evolution. The first profit-oriented wind turbines were installed in2010, and it is predicted that the overall potential in wind turbine energy generation by 2016 would be 127MW. The very first solar power plants that were used commercially were dispatched in 2016, the potential of profit-oriented solar energy plants by the end of 2016 was 21MW, further 50MW of rooftop solar plants were additionally associated by the end of 2016. The solar power plants at a smaller scale have started dissipating advancements, and possibility examines was started to work on the concept of solar power plants in the park. In the marshy lands found in the north of Colombo, a small installation of a solar plant is being located. (*Plan for Long-Term Generation Expansion 2018-2037, 2018).*

1.1.4 Electricity Sales and Demand Growth

CEB broad casting individual as the assigned one buyer also broad casting service provider. There are six licenses held by CEB: a license for generation, a second one for transmission and the bulk supply, and the other four licenses are for distribution. But all of these authorized licenses that CEB holds lack any autonomous proprietorship shape and control.

There are several main tariff categories in Sri Lanka. They are Domestic, Religious, Industrial, General Purpose, Government, Hotel, and Other. At the end of 2018 - 6.3 million customers are provided by the power efficacy of Sri Lanka, and the overall need for country electricity has increased to 14,091GWh, which was only 9,268 GWh before 10 years. The rate of growth of average demand is about 2.6% per year. The highest sales that were found in 2018 were 2,616 MW, and ten years before it was 1,842 MW. The Domestic Religious have contributed the total electricity sales of about 14,091GWh, Industrial, basic objective Hotel Government and Other with 33%, 31%, 24%, and 12% respectively (*Ceylon Electricity Board SALES AND GENERATION DATA BOOK 2018*, 2019).

Sri Lanka's electricity demand was developing at a medium rate yearly that was about 5%-6% throughout the last twenty years, and it is anticipated that this pattern will proceed later on. Effective planning of electricity needs exact estimates of future demand to balance the supply and requirement of electricity; therefore, electricity demand foresee is essential for both power system operation and groundwork. The capacity to precisely anticipates what is to come is critical to numerous choice cycles, such as arranging, booking, buying, technique definition, and strategy making. Along these lines, people consistently attempt to discover accurate forecasting models.

1.2 Motivation

The significant motivation factor of this work is to discover the best forecasting technique for forecast long-term industrial electricity sales in Sri Lanka.

In Sri Lanka, the industrial zone is the binding force of economic growth and the country has had a rapid development in the economy for the past decade. Since the strategic planners make strategic decisions based on these forecasts, the establishment of an accurate and reliable forecasting model for electricity demand, which could give important information for government and the decision-makers of the electricity sector to form strategies and plans of power, is vital for the management of electricity system in Sri Lanka. If a foresee misjudge command, the company might not have the sufficient space and means to meet the necessities of the consumers. If a prediction exaggerates demands, the company suffers the cost of extra scope and spoils. Since neither of these consequences is absolute, it is essential to fix an error-free long-term foresee model.

1.3 Statement of the problem

Power is not a product that can be conserved in a system where demand and supply must be constantly assessed. As a result, a country's power generation must be precisely equal to its consumption. If the overall demand is higher than the generation capability of electricity, it will have a negative impact on all consumers by causing a massive economic burden. Additionally, if excessive electricity is generated that is the actual demand, it will also be the reason for economic loss, and the power plant will be underutilized than its capacity. Therefore, exact forecasting of the demand for electricity in the future is crucial for the planning and development of new electricity generation funding to keep a balance between demand and supply.

The complete quantity of power devoured by individuals has to be offset with the measure of the power produced. It is not possible to store bulk power energy proficiently. To keep this balance between creation and utilization, one should be able to forecast future needs. This research will compare diverse forecasting methods to tracking down the best suitable forecasting method and what are the possible ways to get it registered in the Electrical Power System of Sri Lanka.

Industrial sectors have an essential role in the productive growth of countries like Sri Lanka. The industrial sector is the driving force for economic growth. Past few decades, the no of industrial consumers number increased rapidly in Sri Lanka. Therefore, it is essential to forecast the electricity sales of the industrial sector.

There are so many techniques that have been applied for forecasting the sales of electricity in Sri Lanka, such as Time Series, Neural Network, Machine Learning. But the problem is what are the most suitable predictive techniques to forecast the Electricity sales of the industrial sector in Sri Lanka.?

Therefore, this research focuses on selecting the most suitable predictive techniques to forecast the Electricity sales of the industrial sector in Sri Lanka.

After finding the best prediction technique for forecast the electricity sales in the industrial sector in Sri Lanka with proper analysis, this study will suggest the most appropriate technique for predicting the sales of electricity in the Sri Lankan industry.

1.4 Research Aims and Objectives

1.4.1 Aim

It is necessary to find the best prediction technique and forecast the electricity sales in the industrial sector of the country by analyzing the historical data. Furthermore, it is important to explore the studies and research conducted to estimate the future sales of electricity in Sri Lanka to identify the techniques they have used for electricity sales prediction in the country. This will provide an insight into the gap of the explored areas and the findings/drawbacks of the similar studies conducted. This knowledge will be used to find the best forecasting technique to forecast electricity sales in the Sri Lankan industries.

1.4.2 Objectives

Main goals of this research are:

1. Identify the patterns of electricity sales within the industrial sector in Sri Lanka.
2. Compare the models in terms of forecasting performance.

1.5 Scope

This study will emphasize on long-term requirements of electricity. In the long run, foresees usually are needed for a foresee view of five to twenty years. They are crucial for the long-term planning of capacity needs (system generation and channeling growth), trade, strategical and monetary plans.

As a developing country, Sri Lanka is having increasing electricity demand over time, especially in the industrial sector. Electricity is one of the most important variables influencing economic growth.

Because electricity is required directly or indirectly by all industries, it may be recognized as the backbone of a nation's process. Therefore, the government needs to provide electricity demand from the industrial sector. Estimates of power deals would be critical to electric utilities when choosing development and investments. Therefore, this study will focus on the Industrial sector. The types of industrial customers as the consumers in the industrial sector are as follows,

- Small Industries
- Medium Industries
- Large Industries

Above mentioned customer categories will be considered whenever the analysis is performed.

Energy generation in Sri Lanka is predominantly met by hydro-power. Now energy generates also from coal, thermal and wind. Nevertheless, still, there is a considerable gap between the demand. So, it is imperative to identify the sales in advance, especially for the industrial sector. Past few years, electricity power grids collapsed several times, and it affected the ordinary lives of citizens and the Industrial sector enormously. To stop this kind of situation, it is imperative to be prepared first. To achieve this, analysis and forecast electricity demand is significant.

There are so many techniques that have been using to forecast electricity sales in Sri Lanka. Traditional methods such as Delphi and panel consensus and numerical means such as relapse, episodes, econometric models, and the distribution guide. This is because of specialized evolution and the conventional prediction models; these methods have been extensively concerned to get accurate forecasts. Studying machines has shown positive developments in prediction, it’s quite worthy to explore the uses of studying such types of machines to predict in future. Time Series forecasting techniques and Machine Learning techniques will be considered in the study, and these strategies will be contrasted with one other in terms of the excellence in predicting outcomes.

Also, the scope of this research will take 1969 to 2018 43 years of annual industrial electricity sales data and find the best predictive technique for forecasting the electricity sales of the industrial sector in Sri Lanka.

1.6 Structure of the Thesis

The paper's organization is as per the following: Chapter 2, Background and Literature Review, introduces a literature review of demand predicting applications and identifies the prime kinds of models used in the analysis.

Chapter 3, Methodology, describes the design of the proposing solution, and the methods and models used in this study with all the used tools.

Chapter 4, Results and Evaluation of the paper will give the results of the study. This chapter will also provide a critical discussion evaluating the results.

Chapter 5, Conclusion and Future Works, will wind up the work, showing an outline of the review results and talk about the likely future works.

CHAPTER 2
LITERATURE REVIEW

2.1 Background

There are two basic forecasting methods according to Soliman and Al-kandari, Abu-Shikhah and Elkarmi (Soliman and Al-kandari, 2010) (Abu-Shikhah and Elkarmi, 2011) studies, and they are qualitative and quantitative methods and choosing the suitable mainly depends on available figures.

In qualitative forecasting methods, the future demand is forecasting independently formed on using the ideas of the skilled persons; though they are not only assumptions, yet they are trying to get solid approaches to get good predictions. Thus, these types of techniques are helpful and executed when recorded factual figures are not obtainable. These techniques embrace biased arc fitting, the Delphi method, and technological comparisons.

On the second hand, quantitative forecasting techniques depend on mathematical and statistical formulations. They are highly concerned when the figures/data are obtainable; however, two conditions should be fulfilled: arithmetical information is obtainable from the past; thus, it will be easy to continue this in the future. The quantitative forecasting techniques include an extensive scope of techniques, and every strategy has its characteristics, efficiencies, and prices that should be supposed when selecting a certain method inside certain orders for certain objectives. Quantitative methods involve, relapse study, break down methods, aggressive flow, and the Box-Jenkins method.

Most quantitative forecast issues are time-series data, which needs to collect at a uniform duration from random sampling. As demonstrated in figure 2.1. there are several methods of grouping data and forecast future values.

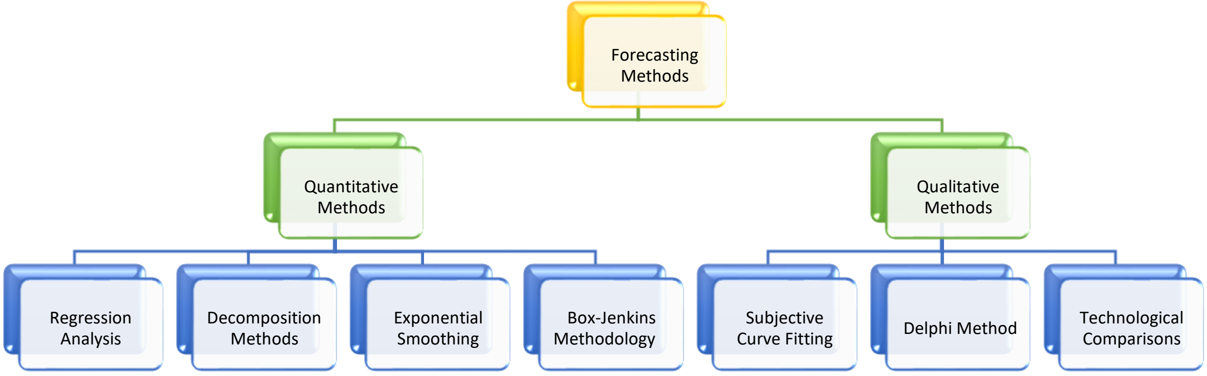


Figure 2. 1 Structure of forecasting method

2.2 A Literature Review

The early electricity load predicting models were predominantly bounded to classical analytical methods, yet with the advancement of the latest studies, load predicting technologies have been impressively evolved. Now predicting models based on the study of machines is becoming more famous day by day. This part of the paper explains the most frequently used forecasting models, regardless of classical or new models.

2.2.1 Time Series Techniques

To forecast different energy sources in the county, Ediger and Akar used Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA) time series methods to the old data from 1950 to 2003 in a relative approach. The perfection of the fitted model is examined with the mean-square error (MS), which is a proportion of the perfection of the fitted model, and the ARIMA anticipating of the all-out essential energy request gives off an impression of being more solid than the summation of the individual figures (Ediger and Akar, 2007).

The same method has been proven by Allah Ditta Nawaz and Niz Hussain 1n 2017. Autoregressive, Integrated Moving Average (ARIMA) methodology used to predicting power utility and CO_2 emitted in the context of Pakistan. For the energy consumption attribute, the Energy they used was obtained from World Development Indicators WDA. To study the prediction outcome, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are used. And fitted ARIMA (2, 1, 2) for CO_2 ejection and ARIMA (2, 1, 1) for energy usage (Allah Ditta Nawaz, Niaz Hussian Ghumro, 2017).

When Allah Ditta Nawaz and Niz Hussain and Ediger and Akar using one or two-time series methods, Akkurt, Demirel, and Zaim used various time frames are Forecasted by using different time series techniques to forecast gaseous petrol utilizations of Turkey. For example, growing rapidly, winters' prediction, and BoxJenkins methods. These techniques are contrasted with each other regarding the prevalence in forecasting performance. They calculated Monthly Natural Gas Consumption using month-to-month petroleum gas usage data in Turkey. To Forecast yearly sales, they utilized yearly gaseous petrol utilization esteems from 1987 to 2008. The mean absolute percentage error (MAPE) and the mean absolute deviation (MAD) counts were used examine the forecast accuracy. The outcomes uncover that in the annual data set, the dual aggressive flowing models outmatch the further substitute predicting models. Then again, in terms of the month-to-month informational, the BoxJenkins SARIMA model gives the preferred outcomes over the others (Akkurt, Demirel, and Zaim, 2010).

However, in 2006 Soares and Souza tried out forecasting electricity demand for different climate changes. They have proposed a theoretical model which implements generalized long memory (Gegenbauer ARMA) to simulate the load's seasonal behavior. An anticipating exercise against a SARIMA model (the benchmark) is profoundly ideal for their model. Mean Absolute Percentage Error (MAPE) count used to the forecast accuracy of the models (Soares and Souza, 2006).

Liu and Lin (1991) made a comparison between Box-Jenkins ARIMA and move purpose models to forecast electricity needs in the populous area in Taiwan. The monthly usage of gas, temperature, and natural gas price between 1975 and 1988. The study used root mean squared error (RMSE) to use measurement statistics to assess the predictability of the models. To acquire further knowledge from the information, they change their plan into quarterly and annual episodes of the natural gas usage, normal temperature, and average natural gas costs to build these episodes. Their study has discovered that ARIMA model beats by the transfer function models.

Another comparison did between the Autoregressive Integrated Moving Average (ARIMA) and a novel configuration combining an autoregressive AR (1) with a high pass filter model to predict electricity usage by using the monthly standard of electric energy usage for the period from 1970 to 1999 by (Saab, Badr, and Nasr, 2001). For the monthly average of electric energy consumption, they found that the AR (1)/high pass filter model yielded the best forecast for this abnormal arrangement of electrical energy data.

2.2.2 Regression Techniques

To make the framework for everyday usage, overall five algorithms were practiced by (Uher *et al.*, 2015) are internal Polynomial Regression, Neural Net, Gaussian Process, Linear Regression , and Polynomial Regression and they found that the best conclusions were attained with a local polynomial regression algorithm. They conducted the analysis based on hourly power use in the Czech Republic from 2011 to2014, as well as the correctness of the preceding model was resolved by manipulating the Mean Square Error (RMSE).

Some researchers combined time series techniques and regression models to forecast demand. First, for measuring the outside temperature and the speed of wind per hour regression model is used. Then, the weekly flow of heat usage as a societal element is included for improvement of the model accuracy. Then used the Seasonal Autoregressive Integrated Moving Average (SARIMA) model with external elements as a mix to take into account meteorological features and traditional heat consumption statistics based on variables. Fang and Lahdelma in2016 used a class of direct regression models (T, T72, T72h, T168, T168h) and the SARIMA model for their study and used Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Theil's Inequality Coefficient (TIC) measurement statistics to examine the forecast accuracy of the models. They found different weekly flows comparatively good with the direct accuracy of straight regression frameworks. Thinking about the closeness of the info, the convenience, and the excellent perfection, the suggested T168h is the best model. The SARIMA model integrated with linear regression can also be registered to forecast the need for heat from a short-term perspective. In any case, it requires a lot of old data to fit the model; the continuity of the historical data is also required (Fang and Lahdelma, 2016).

In addition, two forecasting approaches were merged. Linear Regression achieved the best result with 98 percent accuracy in a study that compares decision tree, linear regression, and fortuitous forest model for forecasting electricity demand done by camurdan and Ganiz in2017. The dataset utilized in this study comprises many measurements and examinations associated with the electricity market in Turkey from 2011 to 2016 daily. Dataset is gathered from public website pages predominantly from the public authority sites revealing the electric market. Mean Absolute Percentage Error (MAPE), R^2 (R Square), Mean Absolute Error (MAE), and Mean Squared Error (MSE) to assess the dependability of the conclusions, evaluation measures were utilised. (Çamurdan and Ganiz, 2017).

2.2.3 Sri Lankan Studies

Based on the historical data, the Classical decomposition approach, Stochastic approach, and Exponential smoothing approach models were used by Pathberiya and Dias in the 2013 study to investigate the best model for predict electricity sales in Sri Lanka. From 2001 to2009, they used monthly power sales data (in GWh) in Sri Lanka and the cityt of Colombo. The identification set for estimating models was data from 2001 to2008, with monthly sales data from 2009 utilised to confirm the models. The study found that the stochastic model provided by ARIMA (0,1,1) (0,1,1)12 produces more perfect forecasts for Colombo and Sri Lanka than the other two models.

Another study Multivariate Regression Approach used to model area wise demand for electricity in Sri Lanka (Ruwanthi and Wickremasinghe, 1999). The analysis used quarterly data for the period 1970 to 1994. 1995 and 1996 data were not included for the analysis due to regular power cuts of the county in these years. Data on the usage of electricity were achieved from the Central bank of Sri Lanka and CEB. The data gathered on kerosene earnings and their sales were taken from Ceylon petroleum and central bank Siri Lanka. GDP at constant factor payments (1980) was utilised as a proxy for consumer income level. Every year, these figures were obtained from the Central Bank of Sri Lanka. They were using Bartlett's Statistic of multivariate regression approach. GDP at constant (1980) factor prices for consumer income level, kerosene price as dependent variable, and electricity demand for their research. The investigation validated the results of several variations on demand for the two periods before and after 1977. Customers' revenue levels are the most powerful on-demand in the post-construction era, while it is the least effective. Prior to1977, it was believed that the influence of pay level had risen significantly as a result of productive initiatives. During the preliberal zed period, the price of electricity had a greater impact on demand. Before1977, the cost of kerosene was a significant factor in determining requirements; after1977, it became less relevant. The value of Bartlett's statistic for the overall importance of the model is (249.79) more significant than the corresponding table value (12.592) at 5% level; it can be assumed that the above regression is highly important; therefore, the variables log (the cost of electricity), log (revenue of customers) acceptably enhance the forecasting of the dependent variables, log (private demand), log(commercial demand), log (business demand).

Using the monthly energy requirements values for the chosen four profitable buildings (Samarawickrama, Hemapala, and Jayasekara, 2016), a computing model is advanced using a regression method called SVMR. In this study, the SVM regression models, utility of future electricity was built for forecasting average electricity. There are many important factors like humidity, average temperature, solar transmissions. Moreover, when this study compared with previous research conducting other approaches to forecast energy usage for the future. Model SVM show maximum forecasting perfection for monthly data.

Madhugeeth and Premaratna, (2008) suggest an ANN solution for electricity demand forecasting. Artificial Neural Networks are examined as a computing model that is able for doing non-linear curve fitting. Three-layered neural network architecture with a backpropagation algorithm is preferred, executed. To evaluate the forecasting results, two techniques were used: RMSE and MAE. The outcome shows that the neural network gives the lowest prediction fault.

Using annual historical data from 1984 to 2015 (Hapuarachchi, Hemapala, and Jayasekara, 2018) proposed another ANN solution for electricity demand forecasting for Sri Lanka. They built a multilayer ANN model for predicting power consumption using data from 1984 to2008, and then validated it using real-world data from 2009 and 2015. Their ANN model performs better on earlier validation data. Their ANN model performs better on earlier validation data.

2.3 Research Gap

According to the literature review, different techniques, namely, regression, stochastic time series autoregressive, ARMA model, ARIMA model, Support Vector Machine based algorithms, and Artificial Neural Networks applied separately to forecasting long term electricity demand in Sri Lanka.

Time-series techniques depend on the beliefs that the data have an inward design, such as autocorrelation, trend, or on and off variation. Time series prediction techniques recognize and study these types of designs. The commonly used time series techniques are Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average with External variables (ARIMAX), and Arima, Autoregressive Moving Average with Exogenous variables (ARMAX).

Support vector machine is a strong method using for resolving groupings and regression issues in the current era. Support vector machines conduct a non-linear chart of the information into an infinite area. At that point, support vector machines use for simple linear functions to produce linear choices limits in the new place.

With the modernization of computing strength, people attempted to resolve the load-forecasting issues using Artificial Neural Networks. Using ANN is to model any complex non-linear relationships, if survive, between the variables that cannot be recognized with old linear models. ANN has to be trained first and checked for its capabilities for generalization.

Even though these techniques are used separately to forecast electricity demand in Sri Lanka, no study compared and analyzed these techniques and suggest the best accurate technique for long-term electricity forecasting, especially in the Industrial sector in Sri Lanka.

In this research, several types of predictive analytic methods, mainly Time Series forecasting techniques and Machine Learning techniques, will be considered in developing the best suitable Long Term electricity demand Forecasting model for the Industrial sector in Sri Lanka.

CHAPTER 3
METHODOLOGY

3.1 Systematic Approach

The systematic approach proposed for electricity forecast models consists of five steps, as shown in Figure 3.1. They are Data gathering and processing, Feature engineering and selection, Built forecasting model, best model selection, and model validation and analysis.

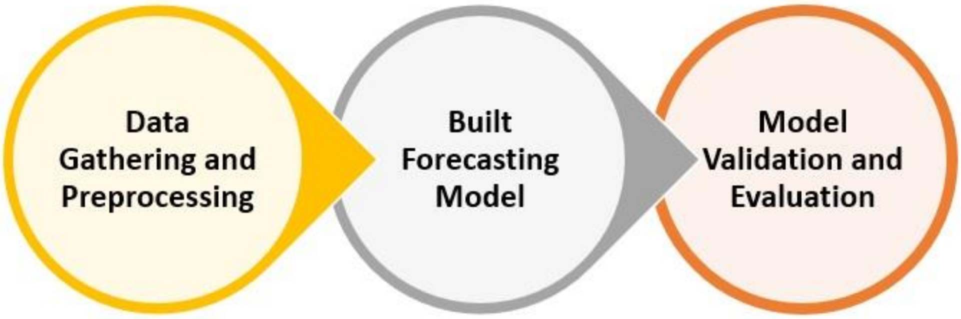

Data Gathering and Preprocessing
Built Forecasting Model
Model Validation and Evaluation

Figure 3. 1 Systematic Approach

Discussion of each phase of the systematic approach listed below.

3.2 Data Gathering and Preprocessing

Study uses the annual historical data of industrial sector electricity demand in Sri Lanka.

Industrial electricity demand data are extracted from the Ceylon Electricity Board, Historical Data Book 1969 – 2018, ad other publications available in their web site

3.3 Built Forecasting Model

3.3.1 Autoregressive Integrated Moving Average (ARIMA)

A famous and extensively used analytical procedure for time series predicting is the Autoregressive Integrated Moving Average (ARIMA) model. This model uses the time series data and numerical data to clarify the data and predict future values. The objective of this model is to elucidate data by using previous value (Investopedia,2019).

This short-form ARIMA is expressive, apprehending the critical sides of the model itself. In short, they are,

The “**AR**" in ARIMA represents autoregression, demonstrating that the model uses the dependent relationship between present data and its previous values. All in all, it shows that the data is regressed on its previous values (Investopedia, 2019).

The “**I**” stands for integrated, which implies that the data is static. Static data applies to time-series data made "static" by eliminating the views from the past values (Investopedia, 2019).

The “**MA**" represents the moving average model, demonstrating that the predictor result of the model relies linearly upon the previous values. Likewise, it implies that the faults in predicting are linear functions of previous faults. Noted, the moving average models are not quite the same as statistical moving averages (Investopedia, 2019).

Every one of the AR, I, and MA elements are involved in the model as a guideline. The parameters are relegated to definite digit values that demonstrate the type of ARIMA model. A standard sign for the ARIMA guideline is appeared and is clarified underneath:

ARIMA *(p, d, q)*

The parameter ***p*** is that the number of autoregressive terms or the number of "lag observations." It is also called the "lag order" and decides the results of the model by giving delayed data points (Investopedia, 2019).

The parameter ***d*** is known as the level of difference. It demonstrates the number of times the lagged signals have been deducted to make the information static (Investopedia, 2019).

The parameter ***q*** is that the number of prediction faults within the model and is additionally alluded to as the size of the movable normal window (Investopedia, 2019).

The parameters take the worth of digits and will be characterized for the model to figure. They will also take a value of 0, inferring that they will not be utilized in the model. In such a manner, the ARIMA model is often transformed to:

- ARMA model (no statistical data, ***d* = 0**)

- AR model (no moving averages or stationary data, just an autoregression on past values, ***d* = 0, *q* = 0**)

- MA model (a moving average model with no autoregression or static data, ***p* = 0, *d* = 0**)

When the parameters (***p, d, q***) have been defied, the ARIMA model plans to appraise the coefficients α and θ, resulting from using past data points to predicting digits.

There are two main methods of time series prediction: univariate and multivariate.

- Univariate uses only the previous values within the statistic to predict future values.
- Multivariable also uses extraneous variables added to the series of values to make the prediction.

The ARIMA model predicts a given statistic supported by its previous values. It is often used for any seasonal need series of numbers that manifest samples and isn't a series of arbitrary occasions. For instance, sales data from a haberdashery would be a statistic because it had been gathered

entire duration of your time. one among the key features is that the data is gathered entire series of continual, uniform breaks. A modified version is often produced to model forecasts over many seasons (Science, 2021). The ARIMA model is becoming a well-liked tool for data scientists to forecast upcoming needs, like sales prediction, making ideas, or stock prices. For example, in prediction stock prices, the model reflects the differences between the values in a series rather than measuring the real values (Science, 2021). ARIMA models can be made in data analytics and data science software like R and Python (Science, 2021).

Advantages of ARIMA Model

The primary benefit of ARIMA predicting is that it needs data on the time series being referred to as it were. To start with, this component is invaluable if one is predicting a large number of time series. Second, this keeps away from a difficulty that sometimes occurs with multivariable models. For instance, consider a model involving salaries, costs, and money. Consistency in money series may be only accessible for a shorter time other than two series., confining the time frame over which the model can be forecast. Third, with multivariate models, the timeliness of data can be complicated. Suppose one creates a large basic model holding only published variables with a long lag, such as wage data. In that case, predictors using this model are absolutely predictions based on predictions of the unattainable observations, adding a source of prediction unreliability (AIDAN MEYLER and GEOFF KENNY, 1998).

Limitations of ARIMA Model

Even though ARIMA models can be exceptionally perfect and valid under the proper circumstances and data accessibility, one of the critical limitations of the model is that the parameters (*p, d, q*) should be manually defined; along these lines, tracking down the most perfect fit can be a long experimentation measure.

Also, the model relies profoundly upon the dependability of old data and the differencing of the data. Guarantee data was gathered precisely for many years, so the model gives exact outcomes and estimates.

ARIMA Model Building

If the data series is static or if there's vast continual that ought to be involved within the model. Seasons are often identified through an autocorrelation plot, a season has a subplot or a spectral plot. The ideas and precisions data will help the info researcher acknowledged the amount of differencing and size of lag that may be required (Box, 2015). The Autocorrelation Function (ACF) is used to figure out the number of MA(*q*) terms within the model. It decides the correlation between the observations at the present point in time and everyone's past points in time. The Partial Autocorrelation Function (PACF) results decide the order of the model or the values for the MA portion of the model. The model order observes how differencing must be used to convert a statistic into a motionless series. The ACF and PACF plots are wont to check extra time faults within the series (Box, 2015).

Augmented Dickey Fuller test (ADF Test)

The Augmented Dickey-Fuller test (ADF test) is a typical measurable test used to test if a given Time series is stationary or not. It is perhaps the most commonly used factual test for determining the stationary of a series. (Selva Prabhakaran, 2019).

To build a Time series model stationary series should change to non-stationary series.

3.3.2 Simple Linear Regression

Linear regression is simply a linear method to model the relationship between the independent variables and the dependent variables. This implies that suppose on the off chance that we have a scatter plot with some points on it, the objective for linear regression is to make a line that can be just about as close as conceivable to every one of the points (Department of Statistics and Data Science, 1997). Sample Linear regression graph is shown in Figure 3.2.

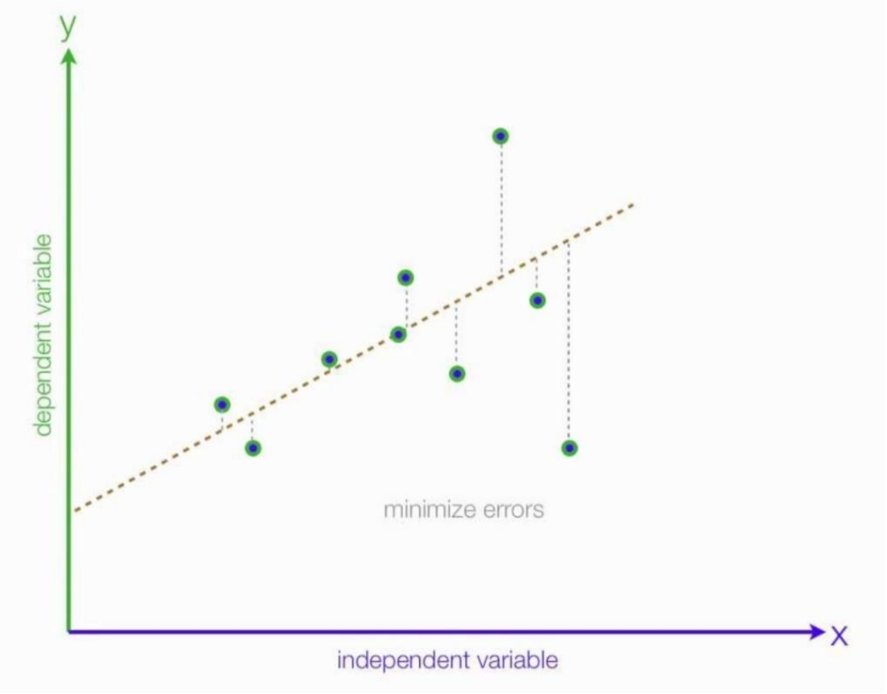


Figure 3. 2 Simple Linear Regression Graph

Before attempting to fit a linear model to observed historical data, a modeller should first determine whether or not there is a link between the variables of interest. This does not imply that one variable causes the other, but rather that the two variables have a crucial connection. A scatterplot can be useful in determining the strength of a link between two variables. If there appears to be no link between the suggested explanatory and dependent variables, then fitting a linear regression model to the data is unlikely to yield a meaningful model. The correlation coefficient is a significant mathematical proportion of interrelationship between two variables, with a value between -1 and 1 reflecting the strength of the link of the observed data for the two variables. (Department of Statistics and Data Science, 1997).

A linear regression line has a condition of the shape Y = a + bX, where X is that the explanatory variable and Y is the dependent variable. The slope or the coefficient of the line is b, and a is that the intercept (the value of y when x = 0).

The most well-known technique for fitting a regression line is the technique of least-squares. The least-squares method finds out the best fitting line for the data by limiting the number of squares of the vertical deviations from every datum to the line. Since the deviations are first squared, then, at that point added, there is no abrogation among positive and negative values.

Advantages of Simple Linear Regression Model

Linear Regression is easy to carry out and simpler to explains the output coefficients.

When you know the relationship between the independent and dependent variables is linear, linear regression is the ideal approach to employ since it is less complicated than other techniques.

Linear Regression is allowing to be over-fitted, yet it tends to be tried not to utilize some capacity lessen methods, regularization (L1 and L2) techniques, and cross-validation(Priyanka Parashar, 2020) .

Disadvantages of Simple Linear Regression Model

Linear regression technique outliers can have vast consequences.

Differently, linear regression expects a straight connection among dependent and explanatory variables. That implies it presumes that there is a straight-line connection between them. It assumes independence between characteristics (Amiya Ranjan Rout, 2020).

3.4 Model Validation

Model verifications refer to the process of confirming that the model achieves its intended goal. In most situations, this will involve confirming that the model is foresighted in the states in which it is being used consciously. This type of validation is accomplished by comparing model reflections to a separate experimental data collection. The data inside in the training set cannot contained in the data collection in the test set.

To validate the Time series Autoregressive Integrated Moving Average (ARIMA) and Linear Regression models for forecast annual industrial electricity sales, hold-out cross-validation is used.

3.4.1 Hold-out Cross Validation

Cross-Validation is a resampling approach that divides the dataset into two sections: training and test data. The train data is used to create the model, while the unseen test data is used to forecast. If the model outperforms the test data and provides high accuracy, it means the model did not overfit the training data and may be utilized for prediction (Lakshana, 2021).

In the hold-out cross-validation method, the original collection of data divide to training and test data sets as shown in Figure 3.3. Typically, the training dataset is more significant than the hold-out dataset. Typical ratios used for split data set include 80:20, 60:40 (Vedas Data, 2018).

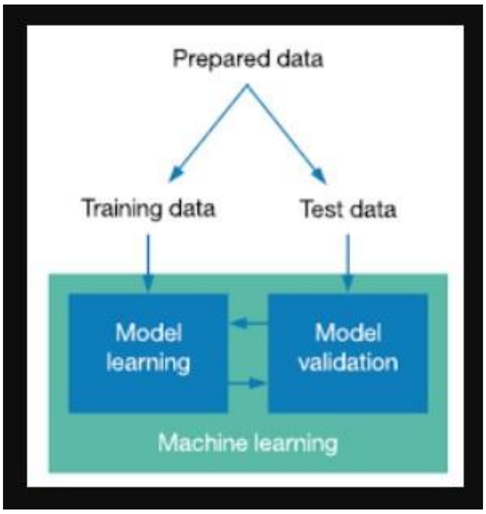


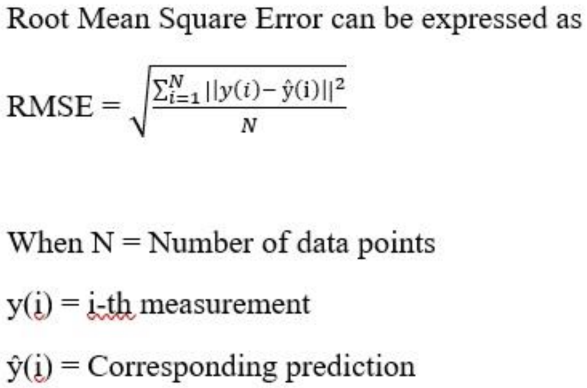
Figure 3. 3Cross validation illustration

3.5 Model Evaluation

After fitting the model, the subsequent stage is to assess the accuracy of that model. Evaluating a model is a vital advance that is performed all through the development of the model. Efficiency Measurement To assess the model performance, the test set was used to measures the real values opposite the forecasted values. Root Mean Square Error (RMSE) and Coefficient of Determination (R^2) techniques were used to calculate the accuracy of the models in this study.

3.5.1 Root Mean Square Error (RMSE)

The Root Mean Square Error (RMSE) is the square root of the difference of the remaining. It shows unquestionably fit of the model to the data–how keenly noticed data points are to the model's forecasted values. As the square root of the difference, RMSE can be explained as the standard deviation of the non-explainable variance and has the valuable property of being in the same units as the response variable. Lower upsides of RMSE demonstrate a better fit. The RMSE is a good indicator of how correctly the model predicts the response. If the major purpose for the model is prediction, this is the primary foundation for fit(Martin Kare, 2013).



Root Mean Square Error can be expressed as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} ||y(i) - \hat{y}(i)||^2}{N}}$$

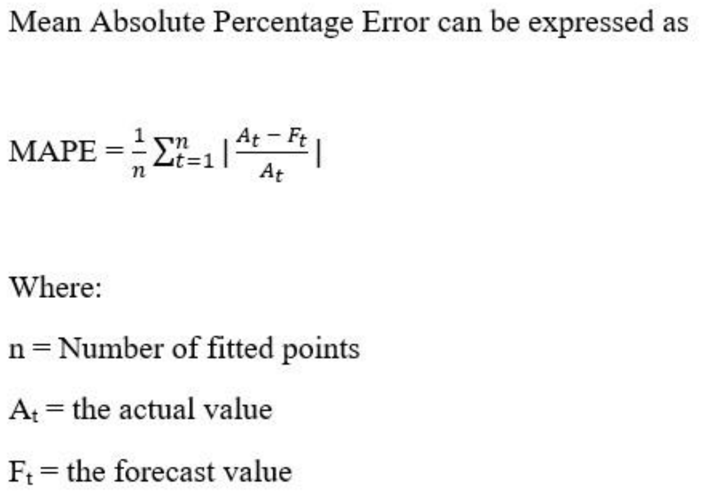
When N = Number of data points

y(i) = i-th measurement

ŷ(i) = Corresponding prediction

3.5.2 Mean Absolute Percentage Error (MAPE)

The Mean Absolute Percentage Error (MAPE) is a proportion of how accurate a predicted model is. It appraises this accuracy as a rate and can be purposeful as the average utter percent fault for each period minus real numbers divided by absolute numbers (Stephanie, 2021).



Mean Absolute Percentage Error can be expressed as

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n} | \frac{A_t - F_t}{A_t} |$$

Where:

n = Number of fitted points

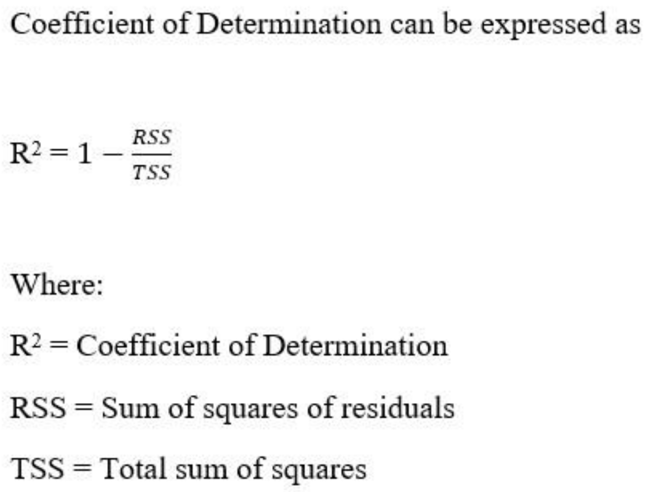
A_t = the actual value

F_t = the forecast value

3.5.3 Coefficient of Determination (R^2)

The coefficient of determination(R^2), also known as "R squared" is a statistical measure that addresses the extent of the difference for the dependent variable simplified by the independent variable.

Though correlation explains the strength of the connection between an independent and dependent variable, R-squared indicates how well one variable's variance describes the variation of another. In this vein, if a model's R2 is0.50, the model's inputs can explain roughly half of the observed variance. (Fernando Jason, 2021).



Coefficient of Determination can be expressed as

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where:

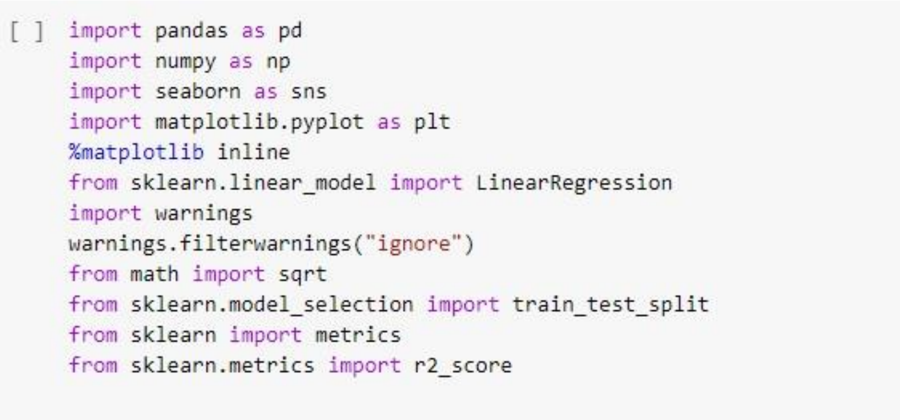
R^2 = Coefficient of Determination

RSS = Sum of squares of residuals

TSS = Total sum of squares

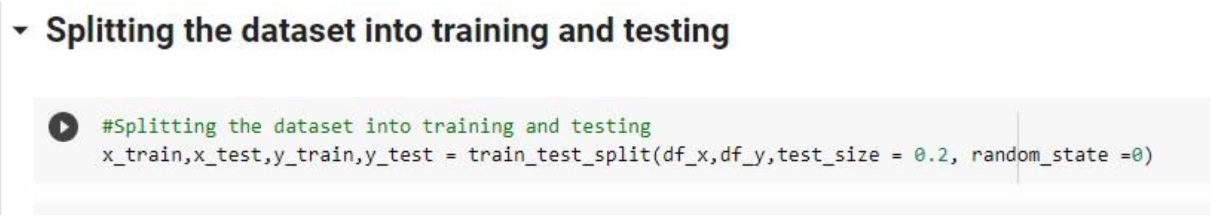
3.6 Machine Learning approaches used for the forecasting models

Linear Regression is a machine learning algorithm based on supervised learning. Using Scikit-learn machine learning library for the Python programming language split the data set in to taking and test data sets for cross validation.



```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.linear_model import LinearRegression
import warnings
warnings.filterwarnings("ignore")
from math import sqrt
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.metrics import r2_score
```

Figure 3. 4 Imported machine learning and statistical packages



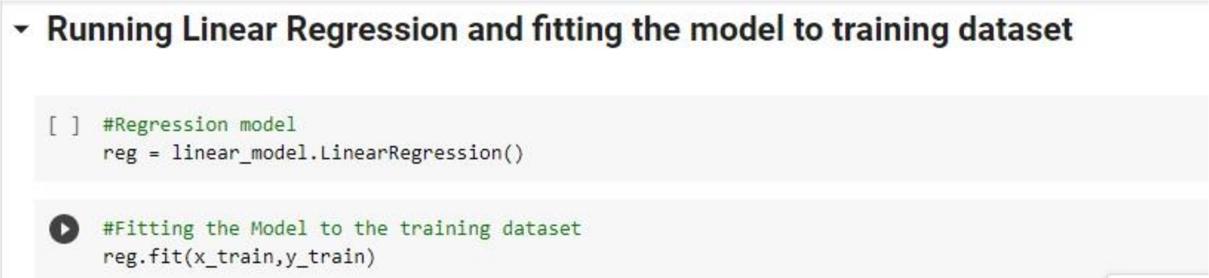
Splitting the dataset into training and testing

```
#Splitting the dataset into training and testing
x_train,x_test,y_train,y_test = train_test_split(df_x,df_y,test_size = 0.2, random_state =0)
```

Figure 3. 5 Splitting the data set for cross validation

Above Figure 3.4 shows the machine learning python packages and statistical packages used for the linear regression forecasting. Figure 3.5 shows how the data set divided in to training and testing data sets for cross validate the implemented model.

Using the same Scikit-learn machine learning library Linear Regression model implemented and fitted to the training dataset as shown in the Figure 3.6.



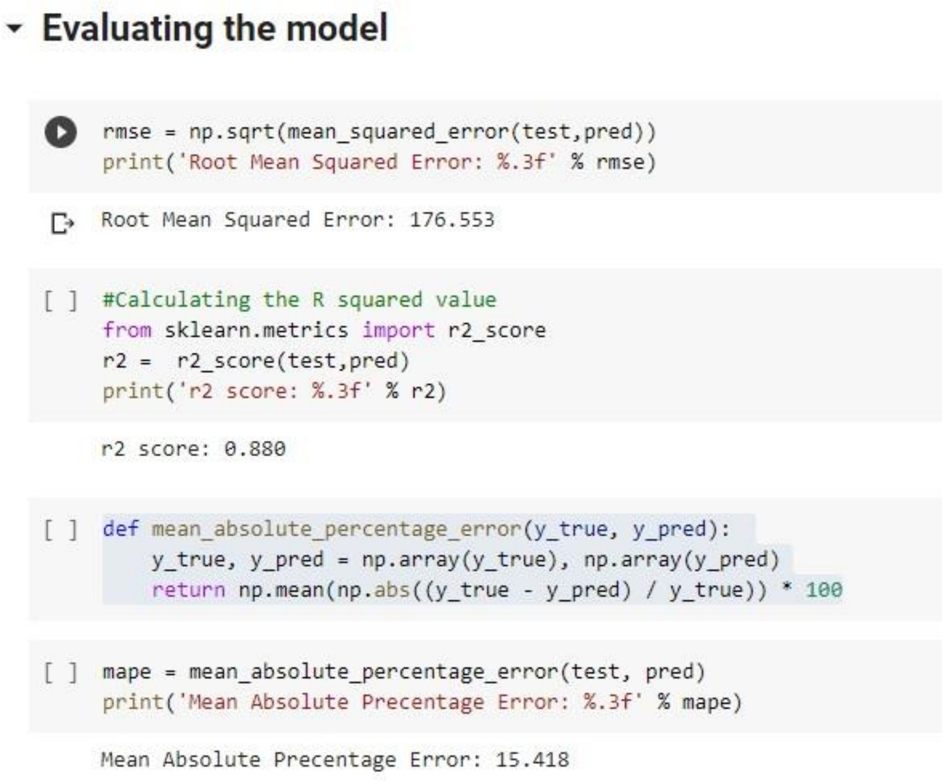
Running Linear Regression and fitting the model to training dataset

```
#Regression model
reg = linear_model.LinearRegression()
```

```
#Fitting the Model to the training dataset
reg.fit(x_train,y_train)
```

Figure 3. 6 Implement the linear regression model

To evaluate both time series ARIMA model and the Linear Regression model used the Scikit-learn machine learning library. Below Figure 3.7 and Figure 3.8 shows the source code how it done.



Evaluating the model

```
rmse = np.sqrt(mean_squared_error(test,pred))
print('Root Mean Squared Error: %.3f' % rmse)
```

```
Root Mean Squared Error: 176.553
```

```
#Calculating the R squared value
from sklearn.metrics import r2_score
r2 =  r2_score(test,pred)
print('r2 score: %.3f' % r2)
```

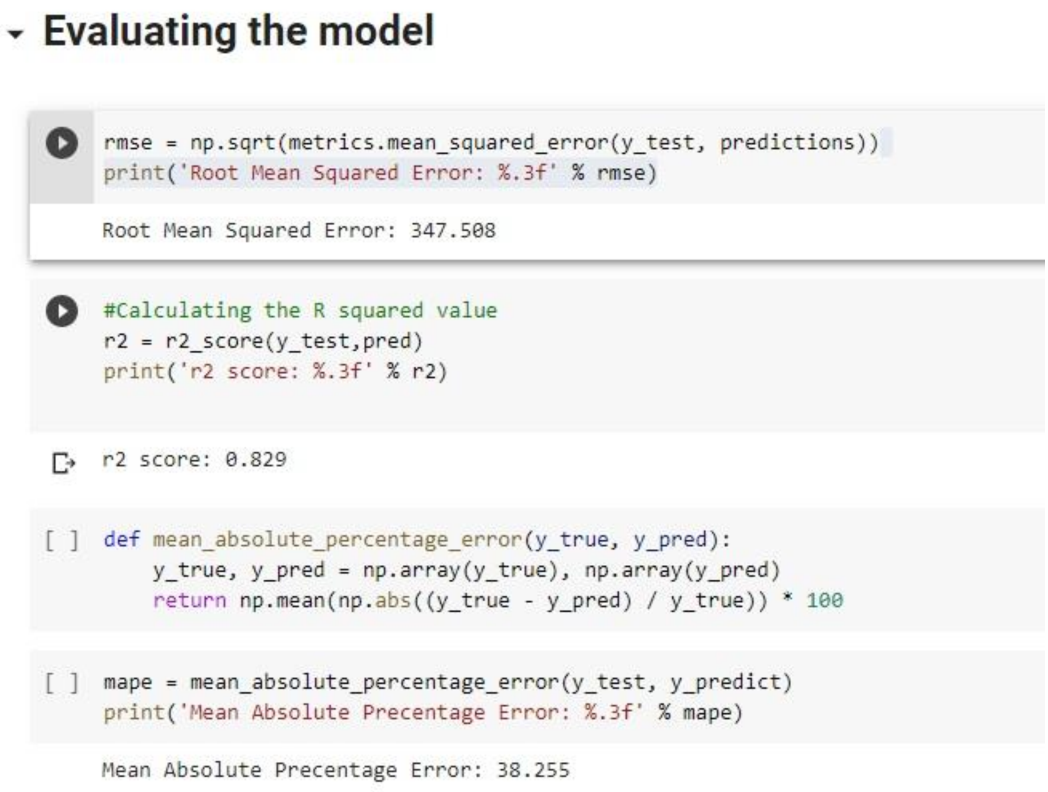
```
r2 score: 0.880
```

```
def mean_absolute_percentage_error(y_true, y_pred):
    y_true, y_pred = np.array(y_true), np.array(y_pred)
    return np.mean(np.abs((y_true - y_pred) / y_true)) * 100
```

```
mape = mean_absolute_percentage_error(test, pred)
print('Mean Absolute Precentage Error: %.3f' % mape)
```

```
Mean Absolute Precentage Error: 15.418
```

Figure 3. 7 ARIMA (0,2,1) model evaluation



Evaluating the model

```
rmse = np.sqrt(metrics.mean_squared_error(y_test, predictions))
print('Root Mean Squared Error: %.3f' % rmse)
```

```
Root Mean Squared Error: 347.508
```

```
#Calculating the R squared value
r2 = r2_score(y_test,pred)
print('r2 score: %.3f' % r2)
```

```
r2 score: 0.829
```

```
def mean_absolute_percentage_error(y_true, y_pred):
    y_true, y_pred = np.array(y_true), np.array(y_pred)
    return np.mean(np.abs((y_true - y_pred) / y_true)) * 100
```

```
mape = mean_absolute_percentage_error(y_test, y_predict)
print('Mean Absolute Precentage Error: %.3f' % mape)
```

```
Mean Absolute Precentage Error: 38.255
```

Figure 3. 8 Linear Regression model evaluation

CHAPTER 4
RESULTS AND EVALUATION

4.1 Data set Analysis

Information used in the estimation consists of annual Industrial electricity demand data over the period 1969 – 2018 inclusive. It would have been interesting had the analysis been done on monthly data instead of annual data, but no reliable monthly data for the period of 1969-2003 is available to the authors.

First, data set checked for the missing values and there were zero columns that have missing values.

4.1.1 Time Plot

As the subsequent stage, time series plots of industrial power utilization were developed and checked whether the data includes a seasonality modification.

As in the below Figure 4.1 there is no seasonality in the annual industrial electricity sales data set in Sri Lanka.

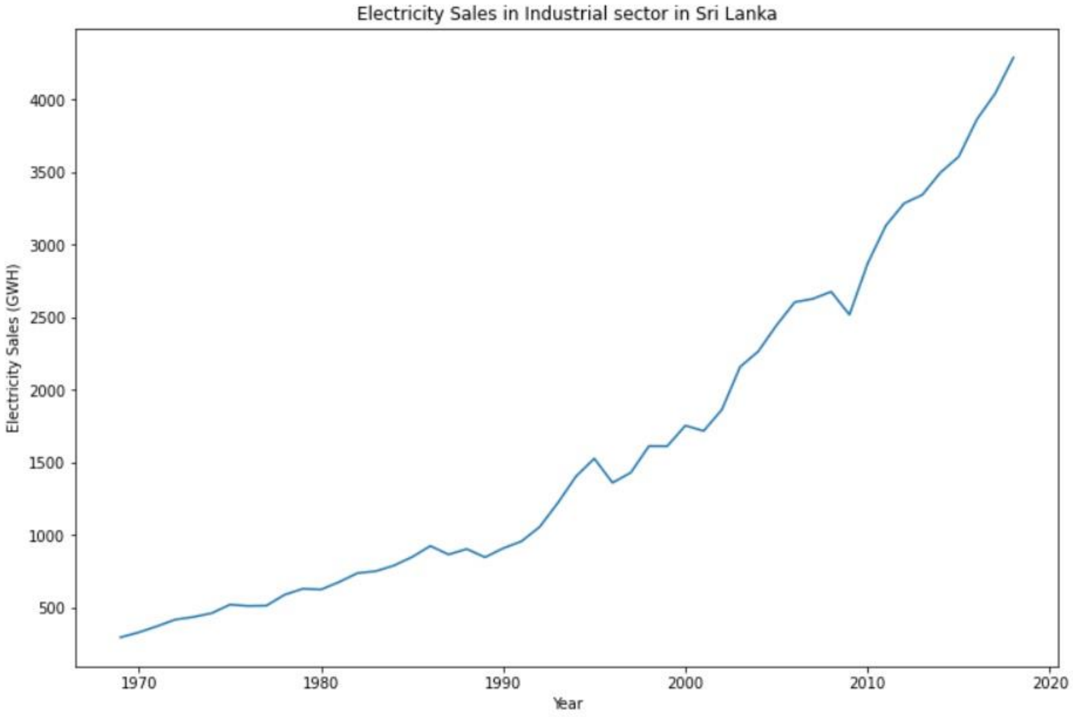


Figure 4. 1 Time plot

As shown in the Figure 4.1 from year 1969 to 1995, industrial electricity sales state a sleek quadratic trend, however when year 1995, it has slight change. Overall annual industrial electricity sales have a rapid trend throughout the past thirty years.

4.2 Autoregressive Integrated Moving Average (ARIMA) Results

4.2.1 Augmented Dickey Fuller test (ADF Test)

The underlying advance performed by utilizing the Unit Root test method named ADF to estimate the stationary conditions of industrial electricity sales. In Table 4.1 assessed results are introduced.

Table 4. 1 Augmented Dickey Fuller test results for original data set

Data set	ADF test	P Value
Industrial electricity Sales	3.590168	1.000000

Because of the original data of industrial electricity utilization and the first differenced data both are not stationary, the second differenced data were considered. In keeping with the Table 4.2 outcomes, second differenced data of industrial electricity sales were stationary under the 0.05 level of significance.

Table 4. 2 ADF test results for 2nd differenced data set

2d Differenced Data set	ADF test	P Value
Industrial electricity Sales	-4.636959	0.000110

Augmented Dickey Fuller (ADF) test results gave a confirmation on stationarity in the annual industrial electricity sales data set in Sri Lanka.

4.2.2 Identifying AR and MA using Autocorrelation (ACF) and Partial Autocorrelation (PACF) Plots

The conduct of the ACF and PACF was inspected deliberately to recognize the potential stochastic models.

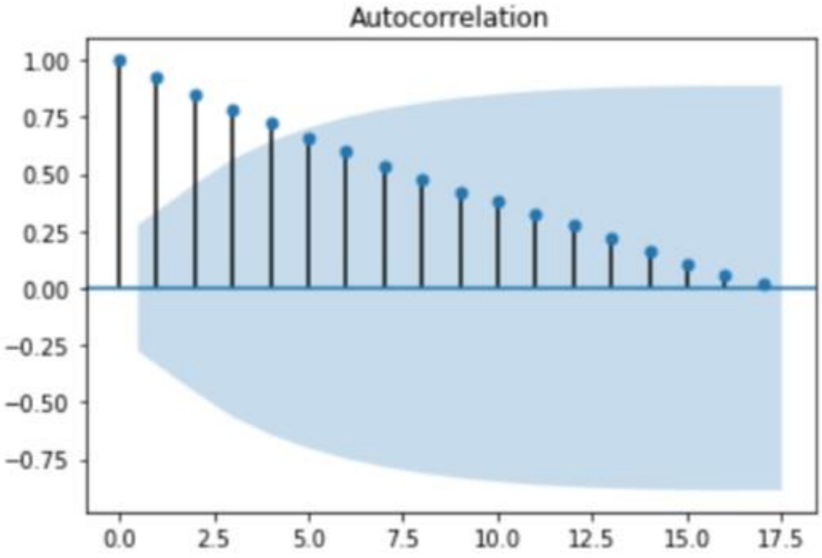


Figure 4. 2 Autocorrelation Plot

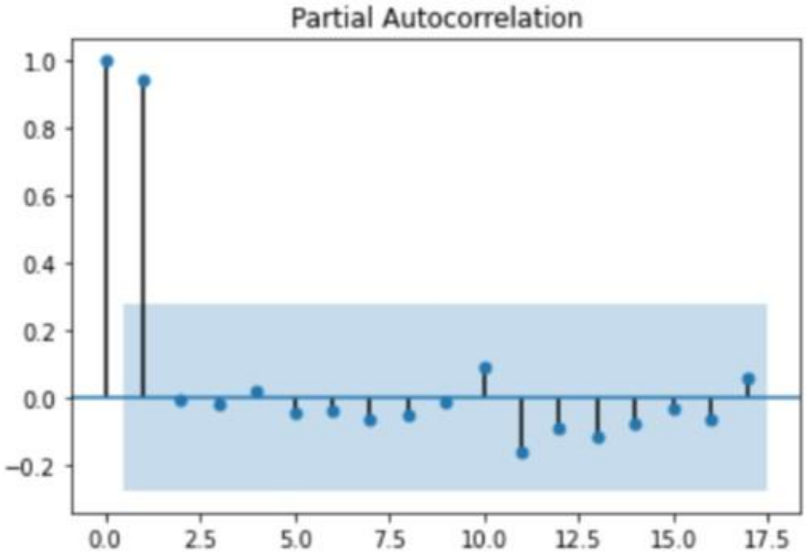
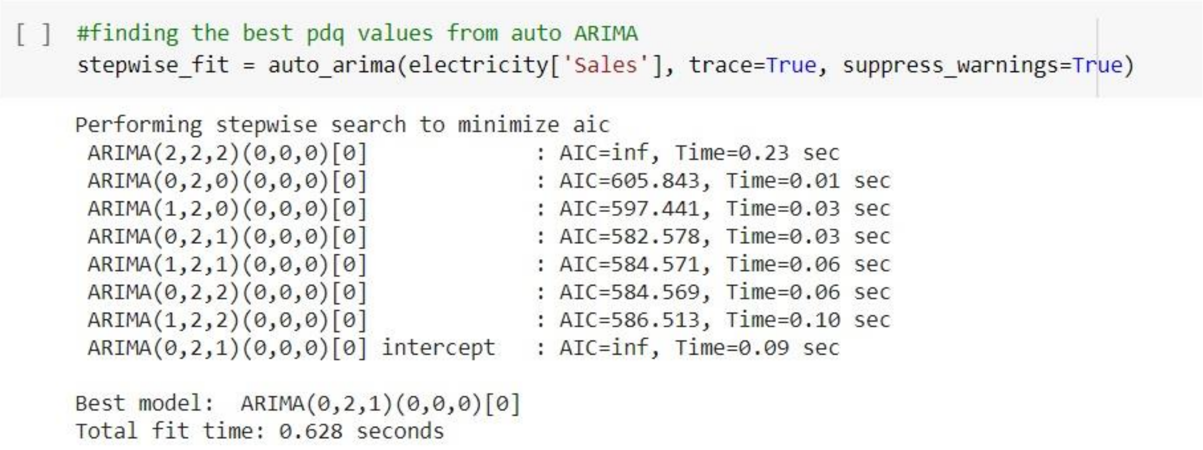


Figure 4. 3 Partial Autocorrelation Plot

As shown in the Figure 4.2 and Figure 4.3, they each offer a reasonable pattern. PACF gradually tightens to 0, even though it has two spikes at lags 1 and 2. On the other hand, the ACF plot shows a tightening design, with slacks gradually corrupting towards 0.

First, tried to fit the ARIMA (2,2,5) model with the values shown in the ACF and PACF for p and q. But it was not successful. Therefore, changed the values of p and q and tried a few different models. Best fitted model is ARIMA (0,2,1) with 582.578 Akaike Information Criteria (AIC) value. As shown in figure 4.4 Same ARIMA model (ARIMA (0,2,1)) gave from the Auto Arima   function for the best fitted ARIMA model for the industrial electricity data set in Sri Lanka.



```
[ ]  #finding the best pdq values from auto ARIMA
     stepwise_fit = auto_arima(electricity['Sales'], trace=True, suppress_warnings=True)

     Performing stepwise search to minimize aic
      ARIMA(2,2,2)(0,0,0)[0]             : AIC=inf, Time=0.23 sec
      ARIMA(0,2,0)(0,0,0)[0]             : AIC=605.843, Time=0.01 sec
      ARIMA(1,2,0)(0,0,0)[0]             : AIC=597.441, Time=0.03 sec
      ARIMA(0,2,1)(0,0,0)[0]             : AIC=582.578, Time=0.03 sec
      ARIMA(1,2,1)(0,0,0)[0]             : AIC=584.571, Time=0.06 sec
      ARIMA(0,2,2)(0,0,0)[0]             : AIC=584.569, Time=0.06 sec
      ARIMA(1,2,2)(0,0,0)[0]             : AIC=586.513, Time=0.10 sec
      ARIMA(0,2,1)(0,0,0)[0] intercept   : AIC=inf, Time=0.09 sec

     Best model:  ARIMA(0,2,1)(0,0,0)[0]
     Total fit time: 0.628 seconds
```

Figure 4. 4 Results of Auto Arima

4.3 Simple Linear Regression Results

Industrial sector electricity data set in Sri Lanka fitted to the linear regression model using sklearn and fitted the model to the training dataset calculate the Coefficient and Intercept shown in the Table 4.3.

Table 4. 3 Calculate the Coefficient and Intercept for the training data set

Data set	Coefficient	Intercept.
Industrial electricity Sales	77.1627	-152191.9297

Below Figure 4.5 shows the linear regression line fitted for the training data set of Industrial electricity data set in Sri Lanka. And Figure 4.6 shows the linear regression line fitted for the testing data set of Industrial electricity data set in Sri Lanka.

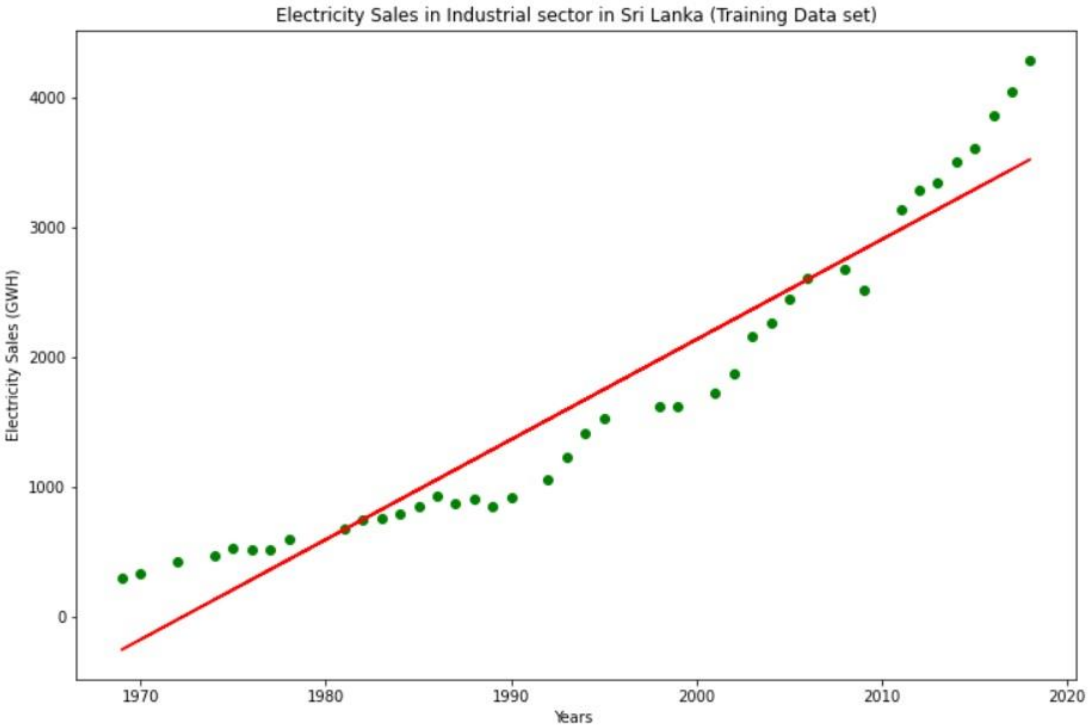


Figure 4. 5 Linear Regression line fitted for the training data set

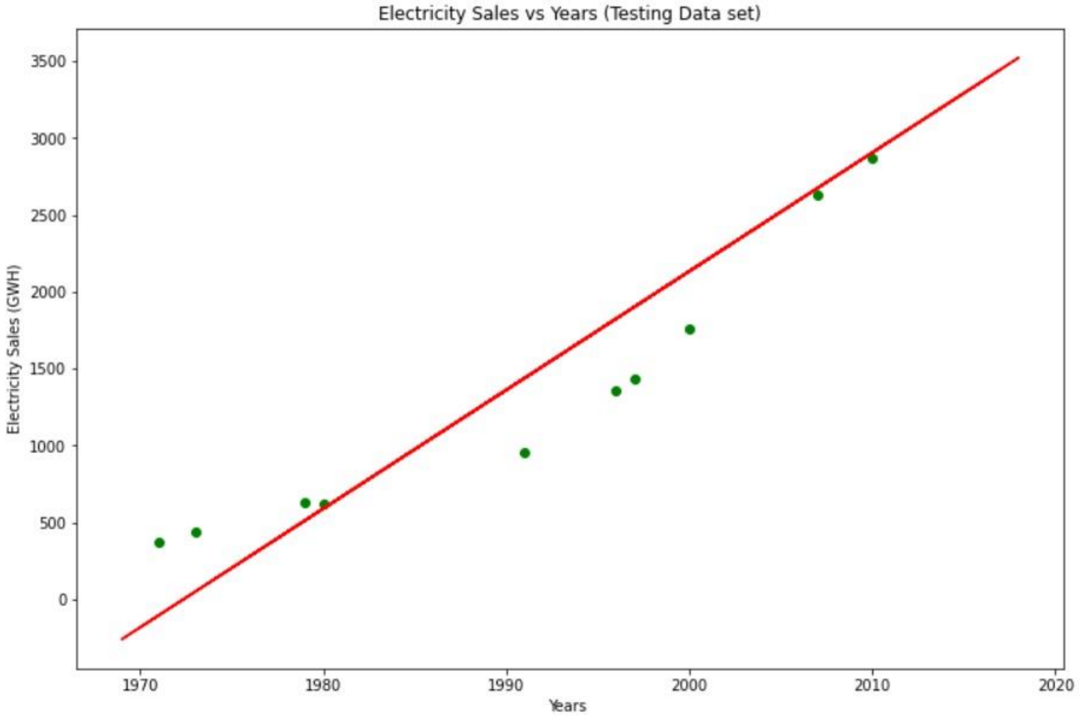
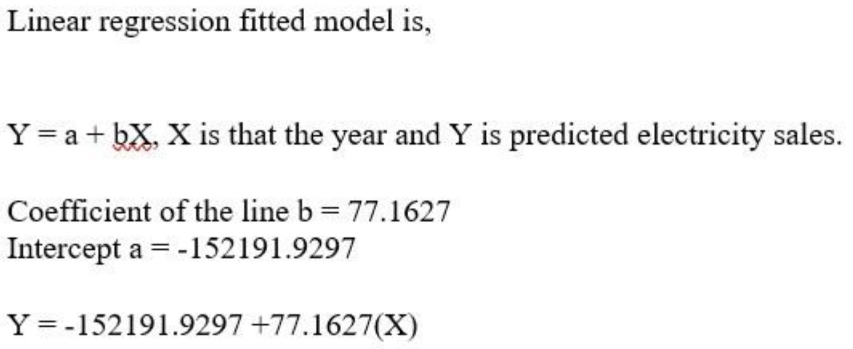


Figure 4. 6 Linear Regression line fitted for the testing data set



Linear regression fitted model is,

Y = a + bX, X is that the year and Y is predicted electricity sales.

Coefficient of the line b = 77.1627
Intercept a = -152191.9297

Y = -152191.9297 +77.1627(X)

4.4 Predicted Values

The created model applied to the calendar years 2009 -2018 and the predicted Industrial sector electricity sales in Sri Lanka for the same period of testing data set from the Time series ARIMA (0,2,1) model and Linear regression model with actual Industrial sector electricity sales are shown in the Table 4.4.

Table 4. 4 predicted Industrial sector electricity sales from the Time series ARIMA (0,2,1) model and Linear regression model with actual Industrial sector electricity sales

Year	Industrial Sector Electricity Sales (GWH)	Predicted Industrial Sector Electricity Sales (GWH) from Time Series ARIMA model	Predicted Industrial Sector Electricity Sales (GWH) from Linear Regression model
2009	2518	2793.07	2828.02
2010	2871	2911.80	2905.18
2011	3132	3033.20	2982.34
2012	3285	3157.28	3059.51
2013	3344	3284.02	3136.67
2014	3498	3413.43	3213.84
2015	3607	3545.52	3291.00
2016	3864	3680.27	3368.16
2017	4042	3817.69	3445.32
2018	4289	3957.78	3522.49

4.5 Evaluating the models

To forecast future industrial sector electricity sales in Sri Lanka the built ARIMA and Linear Regression models should evaluate. For evaluate the models, Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and coefficient of determination(R^2) measures has been calculated separately.

4.5.1 Evaluating the ARIMA (0,2,1) model

The ARIMA (0,2,1) model, with a low RMSE of 176.553 GWH is determined rapidly. For an uncomplicated model this RMSE esteem is a decent outcome, and The results in Table 4.4 convey that the predicted electricity sales are extremely close to the actual industrial electricity sales values. ARIMA (0,2,1) model MAPE is 15.418; on average, that means the forecast is off by 15.42%. And the accuracy of the model is 84.58%. R^2 measures for the ARIMA (0,2,1) model is 0.880 that means the strength of the relationship between the ARIMA (0,2,1) model and the dependent variable is 88%.

4.5.2 Evaluating the Linear Regression model

The linear regression model shows a RMSE of 347.508 GWH. This RMSE value is also a decent outcome for an uncomplicated model. The forecasted values are much of the time lower than the original values, as shown in Table 4.4. Linear Regression model MAPE is 38.255; on average, that means the forecast is off by 38.26%. And the accuracy of the model is 61.74%. R^2 measures for the linear regression model 0.829 implies that the strength of the relationship between the linear regression model and the reliant variable is 82%.

Table 4. 5 RMSE, MAPE and R2 values of ARIMA (0,2,1) model and Linear Regression model

Model	**RMSE**	**MAPE**	**R^2**
ARIMA (0,2,1) model	176.553 GWH	15.418	0.880
Linear Regression model	347.508 GWH	38.255	0.829

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Conclusion

A definitive purpose of this study was to distinguish the best forecasting model from the time series ARIMA model and Linear regression model, which can be utilized to estimate future electricity sales in the industrial sector in Sri Lanka. The model with the least MAPE, least RMSE and highest R^2 is selected as the best forecasting model. According to the Table 4.5 the comparison of the accuracy of the two models showed that the forecasts generated from the ARIMA (0,2,1) is more precise than of the linear regression model.

Accordingly, it was reasoned that the most fitting model to estimate industrial electricity sales in Sri Lanka is the ARIMA (0,2,1) model. The corresponding Root Mean Square Error was 176.553 GWH, Mean Absolute Percentage Error was 15.41%, and the coefficient of determination is 88% respectively.

Data patterns illustrate that the annual industrial electricity sales have quadratic trend during the past 30 years.

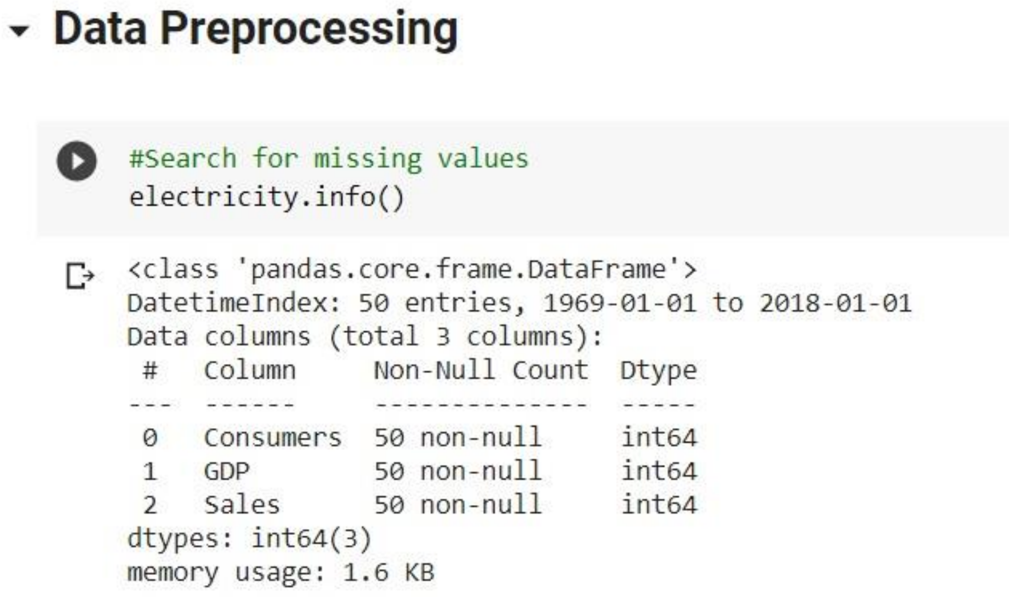
5.2 Future Work

In this research, only the time series ARIMA and Linear regression were utilized. However, many other techniques can be used to predict electricity sales, such as neural network models, data mining techniques, etc. An examination of such practices to recognize methods for the more precise forecast of electricity sales might be helpful. Only the annual electricity sales were used for this study, but it would have been interesting had the analysis been done on monthly data instead of annual data. This study mainly focused on univariate prediction analysis; however, multivariate forecast using other dependent variables would give a more accurate forecast of electricity sales may be helpful.

APPENDICES

Appendices A

Source code for the Time series ARIMA forecasting

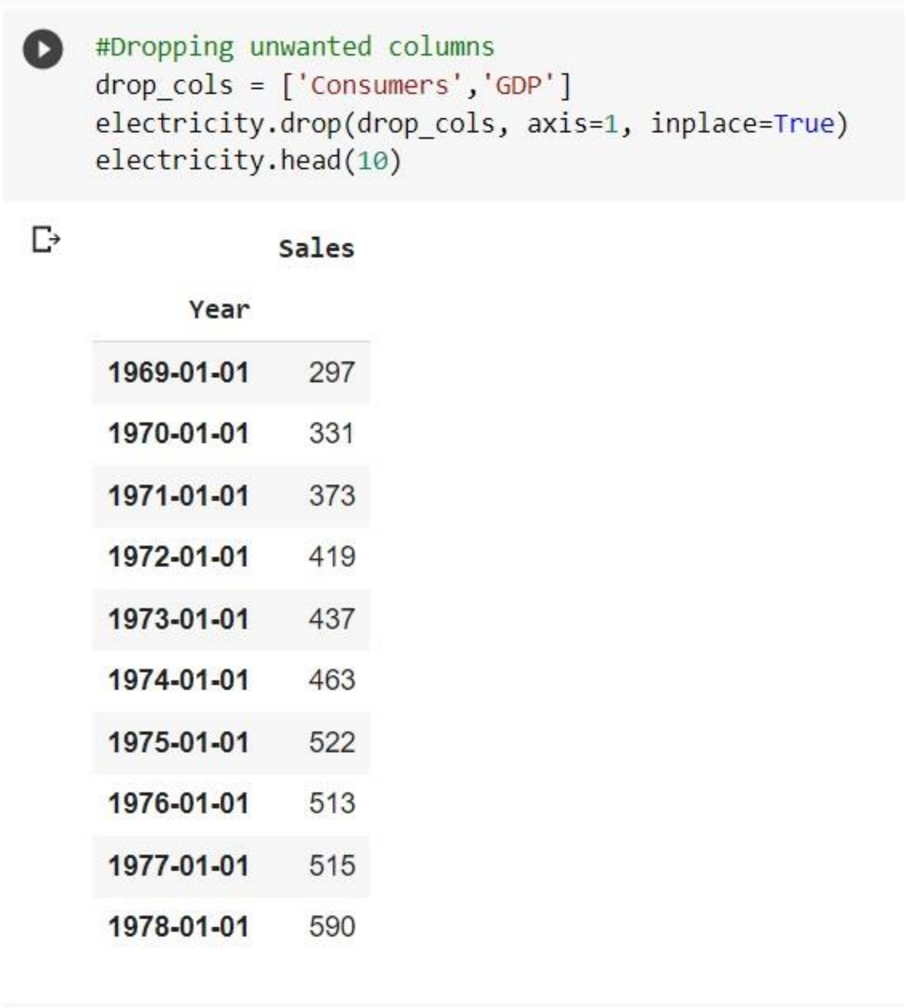


Data Preprocessing

```
#Search for missing values
electricity.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 50 entries, 1969-01-01 to 2018-01-01
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Consumers  50 non-null     int64
 1   GDP        50 non-null     int64
 2   Sales      50 non-null     int64
dtypes: int64(3)
memory usage: 1.6 KB
```

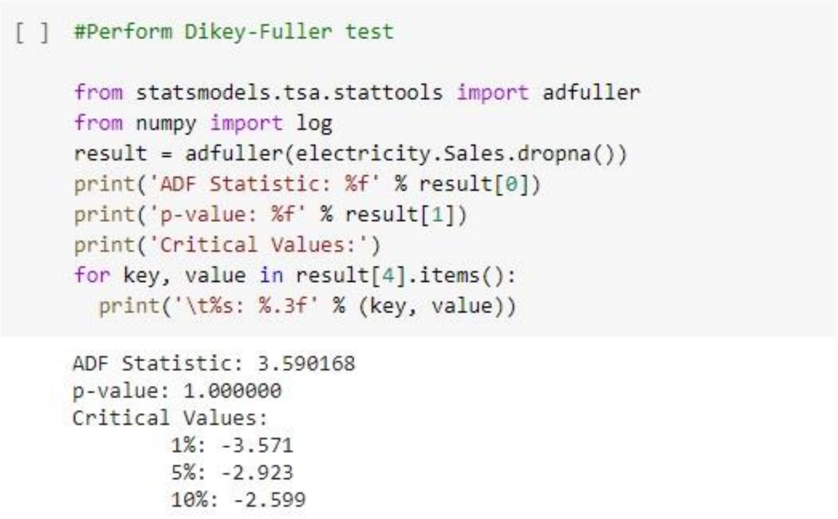
A1. 1 Data Preprocessing for missing values



```
#Dropping unwanted columns
drop_cols = ['Consumers','GDP']
electricity.drop(drop_cols, axis=1, inplace=True)
electricity.head(10)
```

	Sales
Year	
1969-01-01	297
1970-01-01	331
1971-01-01	373
1972-01-01	419
1973-01-01	437
1974-01-01	463
1975-01-01	522
1976-01-01	513
1977-01-01	515
1978-01-01	590

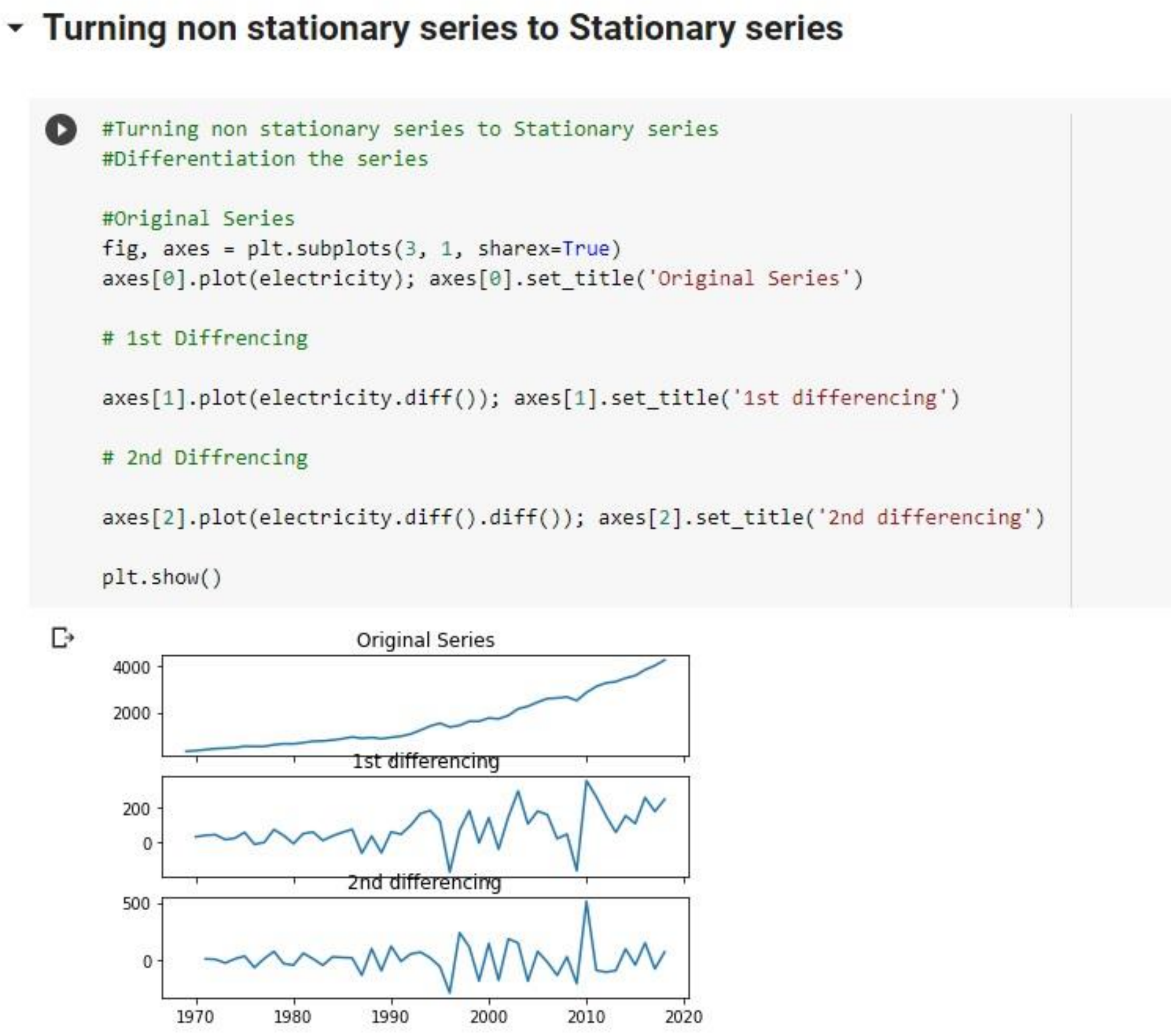
A1. 2 Data Preprocessing: Dropping unwanted columns



```
#Perform Dikey-Fuller test

from statsmodels.tsa.stattools import adfuller
from numpy import log
result = adfuller(electricity.Sales.dropna())
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])
print('Critical Values:')
for key, value in result[4].items():
  print('\t%s: %.3f' % (key, value))
```

```
ADF Statistic: 3.590168
p-value: 1.000000
Critical Values:
        1%: -3.571
        5%: -2.923
        10%: -2.599
```

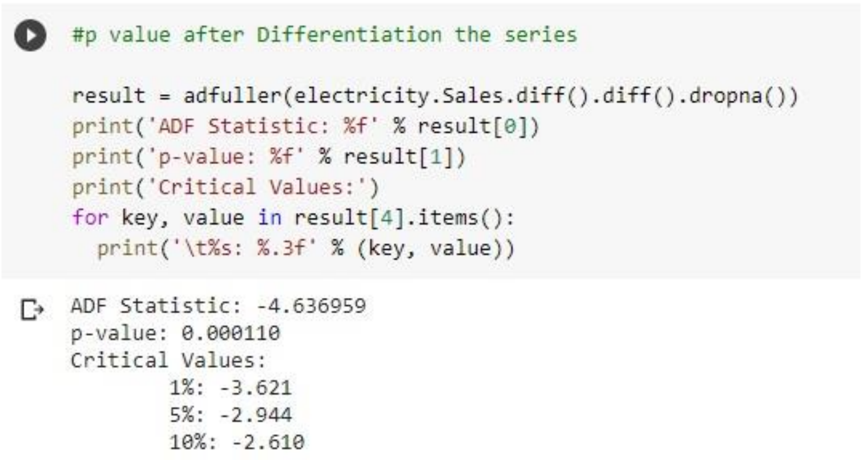
A1. 3 Perform Dickey Fuller test for the original data set



Turning non stationary series to Stationary series

```
#Turning non stationary series to Stationary series
#Differentiation the series

#Original Series
fig, axes = plt.subplots(3, 1, sharex=True)
axes[0].plot(electricity); axes[0].set_title('Original Series')

# 1st Diffrencing

axes[1].plot(electricity.diff()); axes[1].set_title('1st differencing')

# 2nd Diffrencing

axes[2].plot(electricity.diff().diff()); axes[2].set_title('2nd differencing')

plt.show()
```

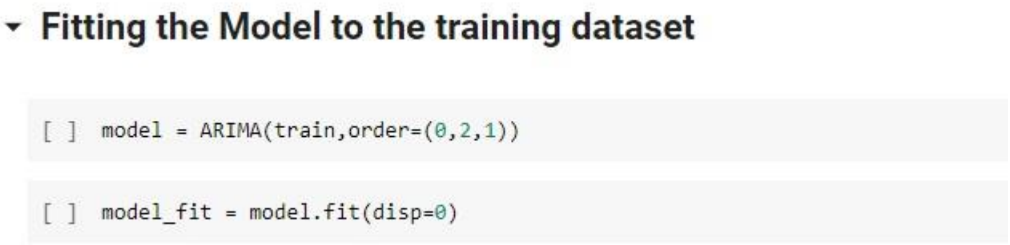
A1. 4 Turning nonstationary series to stationary series



```
#p value after Differentiation the series

result = adfuller(electricity.Sales.diff().diff().dropna())
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])
print('Critical Values:')
for key, value in result[4].items():
  print('\t%s: %.3f' % (key, value))
```

```
ADF Statistic: -4.636959
p-value: 0.000110
Critical Values:
        1%: -3.621
        5%: -2.944
        10%: -2.610
```

A1. 5 Differentiation the series

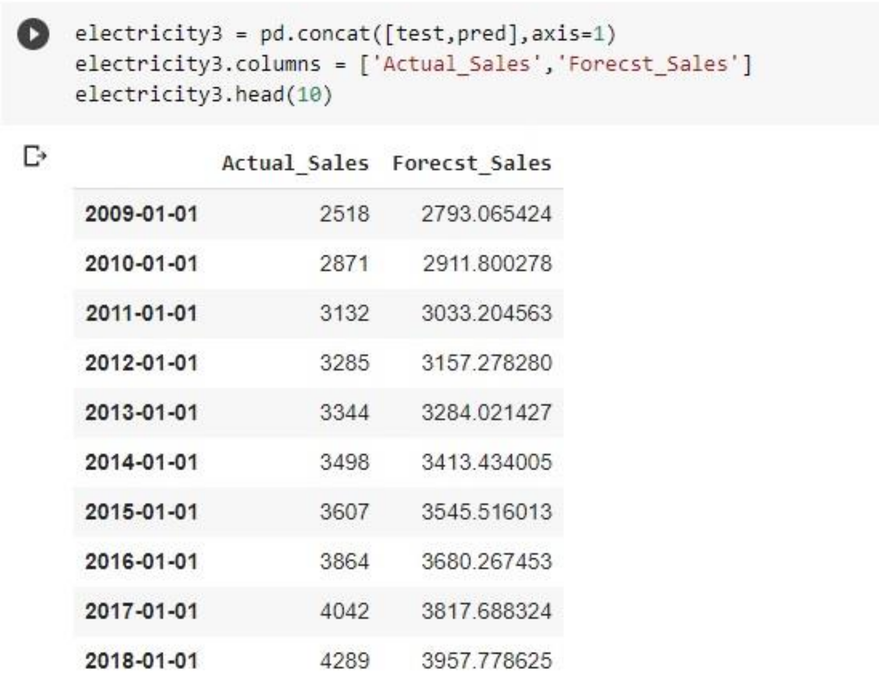


Fitting the Model to the training dataset

```
[ ]  model = ARIMA(train,order=(0,2,1))
```

```
[ ]  model_fit = model.fit(disp=0)
```

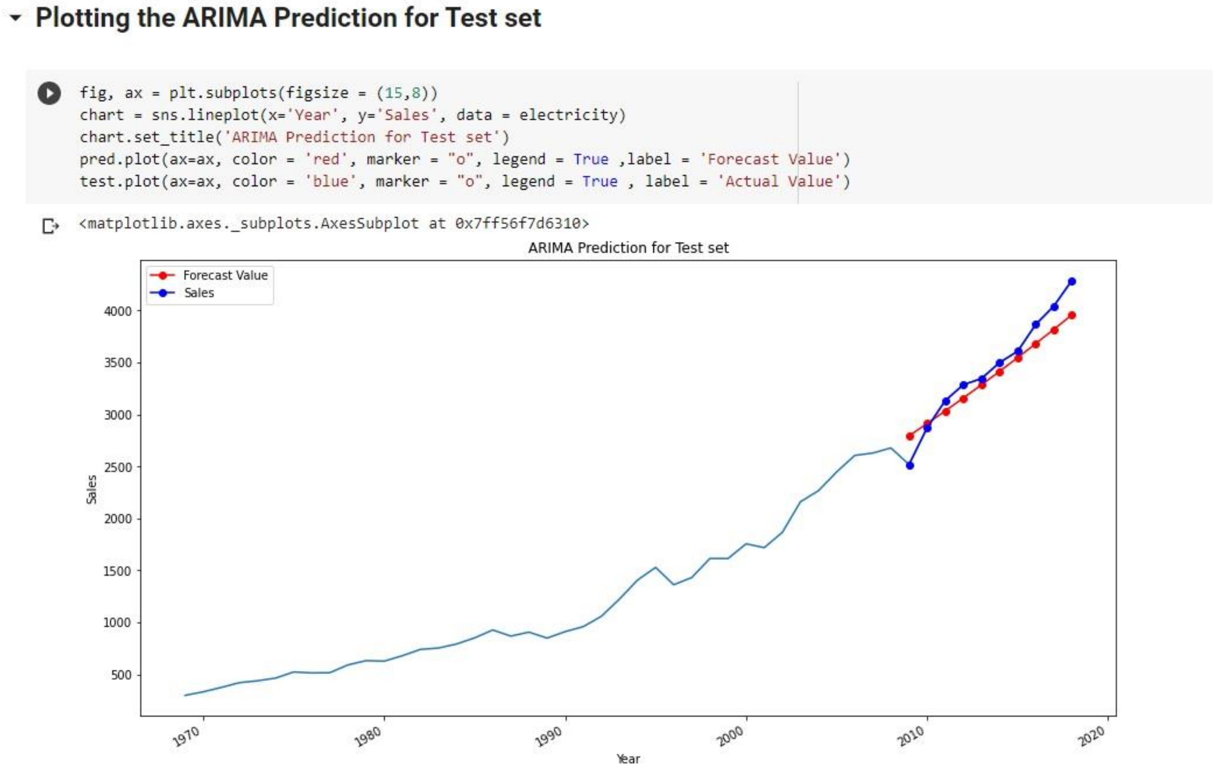
A1. 6 Fitting the model



```
electricity3 = pd.concat([test,pred],axis=1)
electricity3.columns = ['Actual_Sales','Forecst_Sales']
electricity3.head(10)
```

```
            Actual_Sales  Forecst_Sales
2009-01-01          2518    2793.065424
2010-01-01          2871    2911.800278
2011-01-01          3132    3033.204563
2012-01-01          3285    3157.278280
2013-01-01          3344    3284.021427
2014-01-01          3498    3413.434005
2015-01-01          3607    3545.516013
2016-01-01          3864    3680.267453
2017-01-01          4042    3817.688324
2018-01-01          4289    3957.778625
```

A1. 7 Forecast Electricity Sales for test set



▾ Plotting the ARIMA Prediction for Test set

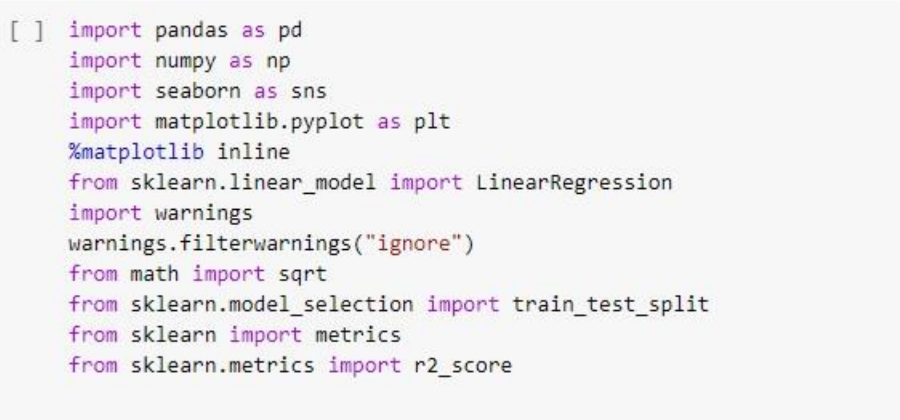
```
fig, ax = plt.subplots(figsize = (15,8))
chart = sns.lineplot(x='Year', y='Sales', data = electricity)
chart.set_title('ARIMA Prediction for Test set')
pred.plot(ax=ax, color = 'red', marker = "o", legend = True ,label = 'Forecast Value')
test.plot(ax=ax, color = 'blue', marker = "o", legend = True , label = 'Actual Value')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff56f7d6310>
```

A1. 8 Plotting the ARIMA(0,2,1) predictions

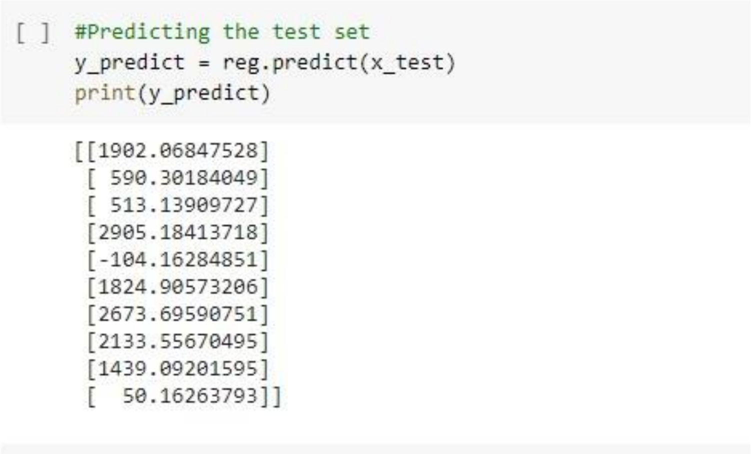
Appendices B

Source code for the Linear Regression forecasting



```
[ ]  import pandas as pd
     import numpy as np
     import seaborn as sns
     import matplotlib.pyplot as plt
     %matplotlib inline
     from sklearn.linear_model import LinearRegression
     import warnings
     warnings.filterwarnings("ignore")
     from math import sqrt
     from sklearn.model_selection import train_test_split
     from sklearn import metrics
     from sklearn.metrics import r2_score
```

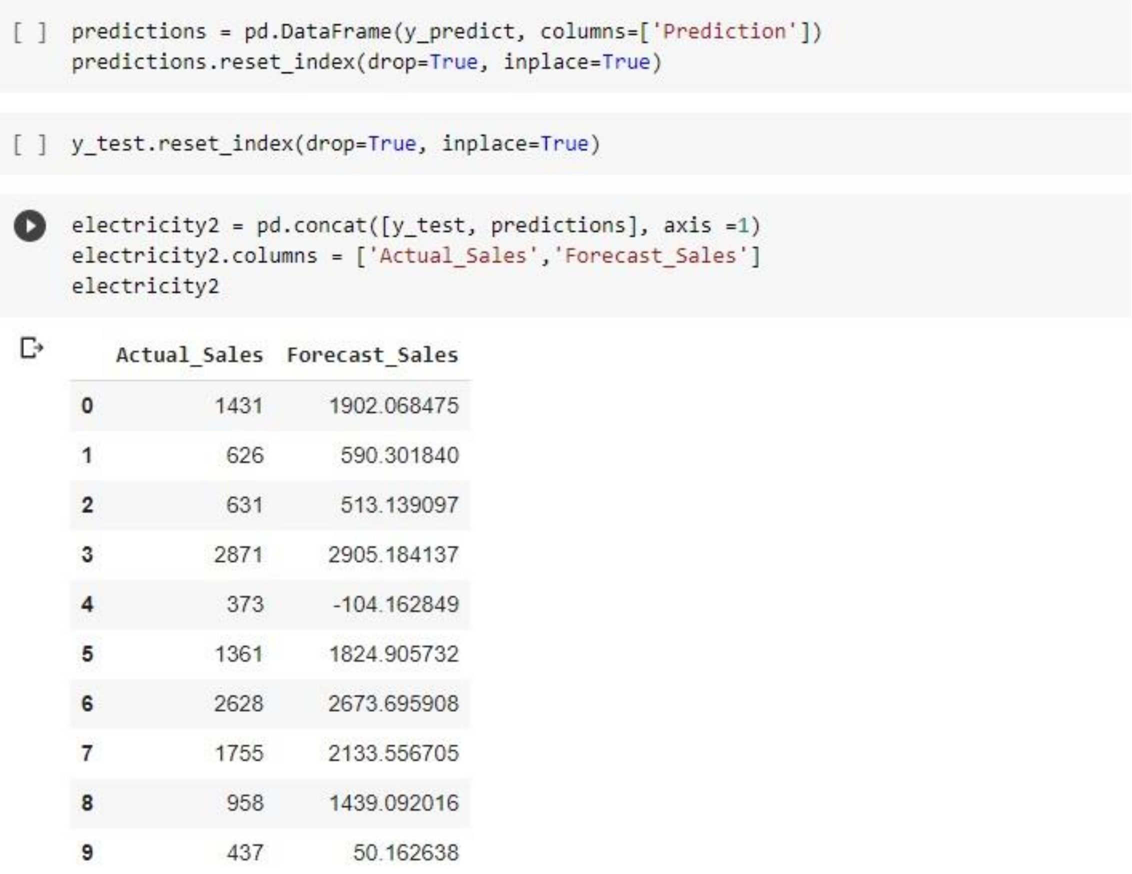
A2. 1Importing Required packages



```
#Predicting the test set
y_predict = reg.predict(x_test)
print(y_predict)
```

```
[[1902.06847528]
 [ 590.30184049]
 [ 513.13909727]
 [2905.18413718]
 [-104.16284851]
 [1824.90573206]
 [2673.69590751]
 [2133.55670495]
 [1439.09201595]
 [  50.16263793]]
```

A2. 2 Predicting the test set



```
predictions = pd.DataFrame(y_predict, columns=['Prediction'])
predictions.reset_index(drop=True, inplace=True)
```

```
y_test.reset_index(drop=True, inplace=True)
```

```
electricity2 = pd.concat([y_test, predictions], axis =1)
electricity2.columns = ['Actual_Sales','Forecast_Sales']
electricity2
```

	Actual_Sales	Forecast_Sales
0	1431	1902.068475
1	626	590.301840
2	631	513.139097
3	2871	2905.184137
4	373	-104.162849
5	1361	1824.905732
6	2628	2673.695908
7	1755	2133.556705
8	958	1439.092016
9	437	50.162638

A2. 3 Forecast Electricity Sales for test set

REFERENCES

Abu-Shikhah, N. and Elkarmi, F. (2011) ‘Medium-term electric load forecasting using singular value decomposition’, *Energy*, 36(7), pp. 4259–4271. doi: 10.1016/j.energy.2011.04.017.

AIDAN MEYLER, A. and GEOFF KENNY, T. Q. (1998) ‘Forecasting irish inflation using ARIMA models’, (11359), pp. 0–8.

Akkurt, M., Demirel, F. and Zaim, S. (2010) ‘Forecasting Turkey’s Natural Gas Consumption by Using Time Series Methods’, *Eur J Econ Political Stud*, 3, pp. 1–21.

Allah Ditta Nawaz, Niaz Hussian Ghumro, and G. M. S. (2017) ‘Forecasting Energy Consumption and CO2 Emission using ARIMA in Pakistan’, *Engineering science and technology international research journal*, 1(January), pp. 53–58.

Amiya Ranjan Rout (2020) *ML - Advantages and Disadvantages of Linear Regression - GeeksforGeeks, Geeks for Geek*. Available at: https://www.geeksforgeeks.org/ml-advantages-and-disadvantages-of-linear-regression/ (Accessed: 12 September 2021).

Box, G. E. P. (2015) *Time series analysis, forecasting and control*. 5th Editio. Edited by G. H. G. David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, R. S. T. Harvey Goldstein, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, and S. Weisberg.

Çamurdan, Z. and Ganiz, M. C. (2017) ‘Machine learning based electricity demand forecasting’, *2nd International Conference on Computer Science and Engineering, UBMK 2017*, (Mi), pp. 412–417. doi: 10.1109/UBMK.2017.8093428.

CEB (no date a). Available at: https://ceb.lk/ceb-history/en (Accessed: 27 February 2021).

CEB (no date b). Available at: https://ceb.lk/corporate-responsibility/en (Accessed: 27 February 2021).

Ceylon Electricity Board SALES AND GENERATION DATA BOOK 2018 (2019). Ceylon Electricity Board. Available at: https://ceb.lk/publication_media/other-publications/81/en. Ediger, V. Ş. and Akar, S. (2007) ‘ARIMA forecasting of primary energy demand by fuel in Turkey’, *Energy Policy*, 35(3), pp. 1701–1708. doi: 10.1016/j.enpol.2006.05.009. Fang, T. and Lahdelma, R. (2016) ‘Evaluation of a multiple linear regression model and

Department of Statistics and Data Science, Y. U. (1997) *Linear Regression, Yale University* . Available at: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm (Accessed: 12 September 2021).

Fernando Jason (2021) *R-Squared Definition, Investopedia*. Available at: https://www.investopedia.com/terms/r/r-squared.asp (Accessed: 12 September 2021).

GDP growth (annual %) - Sri Lanka | Data (no date). Available at: https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?locations=LK (Accessed: 27 February 2021).

GDP per capita (current US$) - Sri Lanka | Data (no date). Available at: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=LK (Accessed: 27 February 2021).

Glossary, M. I. (no date) *R-squared*. Available at: https://www.morningstar.com/InvGlossary/r_squared_definition_what_is.aspx (Accessed: 3 March 2021).

Hapuarachchi, D. C., Hemapala, K. T. M. U. and Jayasekara, A. G. B. P. (2018) 'Long term annual electricity demand forecasting in Sri Lanka by artificial neural networks', *Asia-Pacific Power and Energy Engineering Conference, APPEEC*, 2018-Octob(October), pp. 290–295. doi:
10.1109/APPEEC.2018.8566586.

Investopedia (2019) *Autoregressive Integrated Moving Average (ARIMA), Investopedia*. Available at: https://www.investopedia.com/terms/a/autoregressive-integrated-moving-averagearima.asp (Accessed: 2 March 2021).

Kenton, W. (2021) *Multiple Linear Regression (MLR) Definition, Investopedia.* Available at: https://www.investopedia.com/terms/m/mlr.asp#citation-1 (Accessed: 3 March 2021).
Liu, L. M. and Lin, M. W. (1991) 'Forecasting residential consumption of natural gas using monthly and quarterly time series', *International Journal of Forecasting*, 7(1), pp. 3–16. doi:
10.1016/0169-2070(91)90028-T.

Lakshana (2021) *Cross-Validation Techniques in Machine Learning for Better Model, Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/05/4-ways-to-evaluate-your-machine-learning-model-cross-validation-techniques-with-python-code/ (Accessed: 12 September 2021).

Long Term Generation Expansion Plan 2018-2037 (2018). Available at: https://www.ceb.lk/front_img/img_reports/1532407706CEB_LONG_TERM_GENERATION_E XPANSION_PLAN_2018-2037.pdf.

Madhugeeth, K. P. M. and Premaratna, H. L. (2008) 'Forecasting power demand using artificial neural networks for Sri Lankan electricity power system', *IEEE Region 10 Colloquium and 3rd International Conference on Industrial and Information Systems, ICIIS 2008*, pp. 1–6. doi:
10.1109/ICIINFS.2008.4798394.

Martin Kare (2013) *Assessing the Fit of Regression Models - The Analysis Factor, The Analysis Factor*. Available at: https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/ (Accessed: 12 September 2021).

Priyanka Parashar (2020) *Linear Regression. Linear and Logistic regressions are… | by Priyanka Parashar | Analytics Vidhya | Medium, Medium*. Available at: https://medium.com/analytics-vidhya/various-types-of-linear-regression-937f3c9dda9 (Accessed: 12 September 2021).

Ruwanthi, K. D. R. and Wickremasinghe, W. N. (1999) 'Modelling sector-wise demand for electricity in Sri Lanka using a multivariate regression approach', *Journal of the National Science Foundation of Sri Lanka*, 27(1), p. 55. doi: 10.4038/jnsfsr.v27i1.2977. Saab, S., Badr, E. and Nasr, G. (2001) 'Univariate modeling and forecasting of energy consumption:

The case of electricity in Lebanon', *Energy*, 26(1), pp. 1–14. doi: 10.1016/S03605442(00)00049-9.

Samarawickrama, N. G. I. S., Hemapala, K. T. M. U. and Jayasekara, A. G. B. P. (2016) 'Support Vector Machine Regression for forecasting electricity demand for large commercial buildings by using kernel parameter and storage effect', *2nd International Moratuwa Engineering Research Conference, MERCon 2016*, pp. 162–167. doi: 10.1109/MERCon.2016.7480133.

SARIMA model in forecasting heat demand for district heating system', *Applied Energy*, 179, pp. 544–552. doi: 10.1016/j.apenergy.2016.06.133.

Science, M. in D. (2021) *What Is ARIMA Modeling?* Available at: https://www.mastersindatascience.org/learning/what-is-arima-modeling/ (Accessed: 2 March 2021).

Selva Prabhakaran (2019) *Augmented Dickey-Fuller (ADF) Test - Must Read Guide - ML+*. Available at: https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/ (Accessed: 12 September 2021).

Soares, L. J. and Souza, L. R. (2006) 'Forecasting electricity demand using generalized long memory', *International Journal of Forecasting*, 22(1), pp. 17–28. doi: 10.1016/j.ijforecast.2005.09.004.

Soliman, S. A. and Al-kandari, A. M. (2010) *Electrical Load Forecasting. Modeling and Model Construction, Elsavier Inc*.

Solutions, S. (no date) *Regression - Statistics Solutions, Statistic Solutions*. Available at: https://www.statisticssolutions.com/directory-of-statistical-analyses-regressionanalysis/regression/ (Accessed: 3 March 2021).

Sri Lanka Sustainable Energy Authority (2017) 'Energy Lanka Balance 2017 Sri'.

Stephanie (2021) *Mean Absolute Percentage Error (MAPE) - Statistics How To, Statistics How To*. Available at: https://www.statisticshowto.com/mean-absolute-percentage-error-mape/ (Accessed: 12 September 2021).

Uher, V. *et al.* (2015) 'Forecasting electricity consumption in Czech Republic', *2015 38th International Conference on Telecommunications and Signal Processing, TSP 2015*, pp. 262–265. doi: 10.1109/TSP.2015.7296264.

Vedas Data (2018) *HOLDOUT CROSS-VALIDATION | Data Vedas, Data Vedas*. Available at: https://www.datavedas.com/holdout-cross-validation/ (Accessed: 12 September 2021).

Weedmark, D. (2018) *The Difference Between Bivariate & Multivariate Analyses*. Available at: https://sciencing.com/difference-between-bivariate-multivariate-analyses-8667797.html (Accessed: 3 March 2021).