Predicting solar power generation in USA using timeseries

UMBC

Group – 6 Project report

Submitted by

Karthik Chavali – SD17658

Isha Shah – ZN70176

Maitry Rawal - QT33534

Tejal Ravindra Jadhav - UG05012

Sowmith Chakilela – ZZ04733

TABLE OF CONTENTS

Predicting solar power generation in USA using timeseries

1. ABSTRACT

Climate change has been a major issue for over a century. In the late 1980's a seminar paper was presented by a scientist named (Arrhenius, 2009)' 'On the influence of carbonic acid in the air upon the temperature of the ground mentioned that the carbonic acid is causing the raise in temperatures. As per the article presented by (Valéry Masson, 2014) and others on 'Solar panels reduce both global warming and urban heat island' helps in understanding the usage of solar power and its impact on climate change. We are still trying to solve the issue of what we call it as a man-made disaster. Tackling such issues could help in understand the need and urgency to look for alternate renewable sources.

With machine learning tools we could come up with solution on methods for reducing the greenhouse gas emission and ways to adapt the climate changes by predicting and analyzing the available data. The main goal is to provide an overview of the uses of generating solar power and renewable energy resources with help of machine learning algorithms, The generation of solar power and usage in Maryland, USA and predict how much energy could be produced using solar PV panels based on the previously available production data from U.S. Energy Information Administration. The study also presents with the overview of the problem that we would like to solve, solar power generation system in USA, data sources used in predicting the future production of electricity using solar PV panels, algorithms and methods used to predict the solar generation in Maryland, results obtained.

This report could provide a clear understanding on the usage of how solar energy could help in reducing the greenhouse emissions and reduce the use of non-renewable resources.

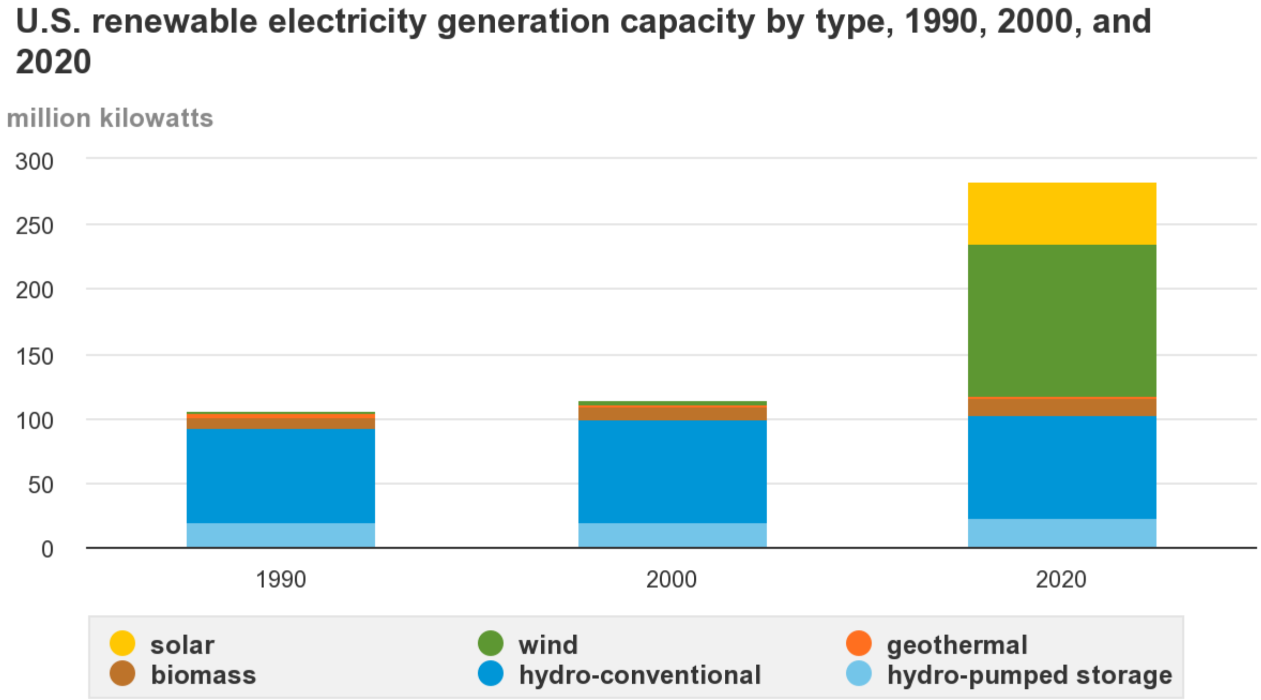
2. INTRODUCTION

2.1.BACKGROUND

In recent times global warming, pollution, greenhouse gasses, natural disasters have been some of the main topics of discussion and has been growing concerns over the usage of sources in generating electricity over the years. The traditional conventional ways of generating electricity have always been an issue around the world and this led to the raise of renewable resources in producing electricity. As per the report published by (*Suparna Ray, 2021*) in US Energy Information systems, *Renewable's account for most new U.S. electricity generating capacity in 2021,* Solar photovoltaics generated 15.4 GW of energy till October 2021. It captures 39% of total electric energy generated leading to top when compared by other sources of electricity generation. EIA forecasts additional 4.1 GW will be entering the service by small scare solar PV by end of this year.

The global change to renewable energy resources has made it easy for the development of solar PV panels and resulting in decrease in cost of producing electricity using solar panels and increase in its efficiency. AS per the (*IRENA,2020*) report *renewable power generation renewable power generation costs in 2020* it states that the levelized cost of electricity has significantly dropped to 85% between 2010 to 2020, that is "USD 0.381/kWh to USD 0.057/kWh and total installing cost fell from 4,731/kW to USD 883/kW".

As the world has come together for the Glasgow climate pact many countries agreed at the UN climate Change Conference (COP26) regarding the phasing out of use of coal and increase the use of more renewable energy resources and become carbon neutral by 2050. Though they had many compromises as mentioned by (UN environment programme,2021) in the article "*COP26 ends with agreement but falls short on climate action*" US has agreed to reduce its emission and make its best efforts to help in limit the global warming to 1.5 degrees.



U.S. renewable electricity generation capacity by type, 1990, 2000, and 2020

Note: Net summer capacity of utility-scale generators. Hydro includes conventional and pumped-storage hydro.
Source: U.S. Energy Information Administration, *Annual Energy Review 2011* and *Electric Power Monthly*, February 2021, preliminary for 2020

Figure 1.The picture represents the usage of renewable and non-renewable resources for past three decades.

This is an important step for all the countries and especially for US to step up and this report provides an insight on the way how these steps are taken in reducing global warming and how it could be achieved. This project uses machine learning algorithms to particularly answer this question **How efficient is solar power helpful in minimizing carbon emission and help in climate change?** As a solution we delve into solving the issue by understanding the uses of renewable resources compared with non – renewable resources and the energy consumption and production over the years in USA. Based on the data collected we would forecast the usage and production of these available resources to come to a conclusion of how efficient and viable it would be to invest in Solar power energy production and meet the agreement of COP 26 by 2050.

2.2. OBJECTIVE

The main objective is to analyze the total production of electricity produced using solar PV panels in the USA. Based on the machine learning algorithms implemented on the datasets collected to predict the further generation of electricity in the state we would estimate the production capabilities by providing an assessment, comparing renewable and non-renewable energy resources.

Apart from the timeseries algorithm we would like to consider other model Regression analysis and compare the results to provide overall accuracy of these models.

2.3. RESEARCH QUESTION

How using renewable resources like solar power for electricity generation could reduce the overall emission of carbon and other greenhouse gases in USA. Along with a forecast of generation of electricity using solar power for over three months using machine learning algorithm.

2.4. LIMITATIONS

With the datasets being collected for the study there are limitations that are considered to focus on the main goal of the project.

We have limited the applications of machine learning algorithms on datasets to three for this study, by considering the previous results yielded in other studies where these algorithms were used on similar studies. To get the desired results for the study we are using timeseries and regression analysis.

The focus would be on applying timeseries algorithm to forecast the data and comparing the results with the other two models. The goal of implementing other two algorithms is to provide a general view of algorithm performances by comparing the results.

3. LITERATURE REVIEW

This project provides the results on the issue of climate change and its investment for further. Applying machine learning algorithms on production/consumption of solar power. Compares data of renewable vs non - renewable resources overall. All data sets were taken from various sources and dependent on Kaggle resources and the methods mentioned below were applied to see how data works and what result it provides with comparison of Mean Absolute Error as this provides the accuracy of an algorithm for timeseries and linear regression model.

3.1.TIMESERIES ALGORITHM

A time series is a collection of statistical data taken at regular intervals. The project derives the implication of timeseries as mentioned the author (W. Yan, 2012) in the article '*Toward Automatic*

Time-Series Forecasting Using Neural Networks' understanding how the time series has behaved in the past relies heavily on the data pattern. The data that is collected for our study is under similar lines and giving us a scope that timeseries algorithm would be more relevant to implement on the datasets of the project.

3.2.LINEAR REGRESSION

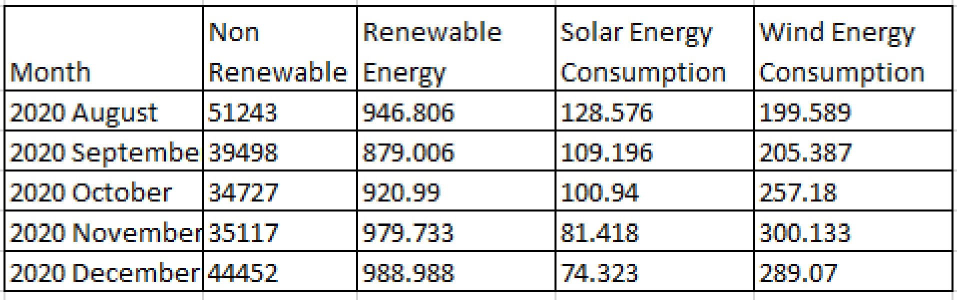
The linear regression model looks for a common relationship between the mean of the independent and dependent variables during classification. As mentioned by the author (*M. Huang, 2020*) in his article '*Theory and Implementation of linear regression*' it assumes that there is a straight relationship between them. However, in some cases, it is very important to scrutinize the dependent variable. This is more difficult and seems to reduce the performance of the model, as the mean values of the variables are not completely sufficient to predict the correct output of the classification without considering the individual variables. Our intent to apply this algorithm for the study is purely for the result comparison and to understand which algorithm provides a better result.

4. DATA PREPARATION

The process used for this project is described in steps on how the data is collected and cleaned and divided based on the training, validation set and test set. Throughout the project we have used WEKA tool to forecast and train the data sets. The data collected for this project are from various sources and based on Kaggle set.

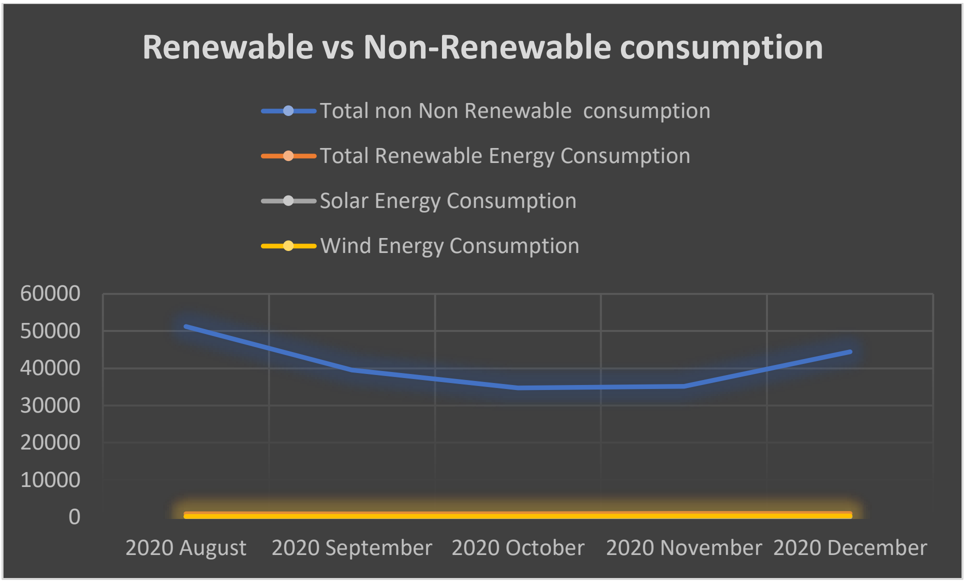
4.1. DATA COLLECTION

The data for the project is collected are based on total energy consumption of both renewable and non-renewable resources along with production of electricity using solar, wind, coal and other various means. Source of these datasets are mainly from US energy information administration. We have collected the data based on the requirements of the project and forecasting. The data of total consumption using coal, solar and wind are in Million Kilowatt-hours. A sample of data collected is represented below where the data of renewable and non-renewable data is shown for a period of five months.



Month	Non Renewable	Renewable Energy	Solar Energy Consumption	Wind Energy Consumption
2020 August	51243	946.806	128.576	199.589
2020 Septembe	39498	879.006	109.196	205.387
2020 October	34727	920.99	100.94	257.18
2020 November	35117	979.733	81.418	300.133
2020 December	44452	988.988	74.323	289.07

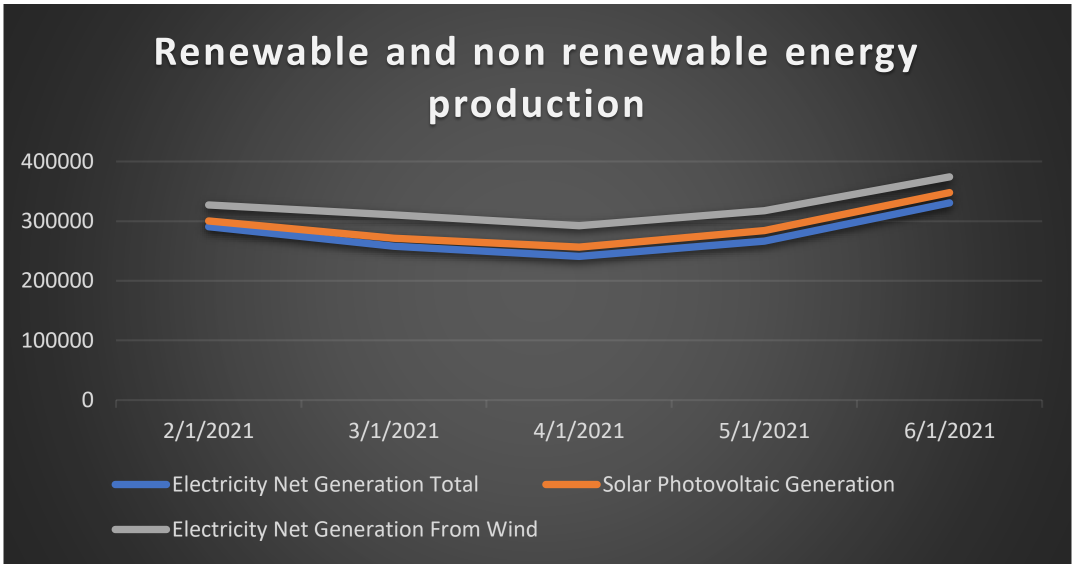
Data of renewable and non – renewable energy consumption



Representation of five months of data of Renewable and non-renewable energy consumption

Month	Electricity Net Generation Total	Solar Photovoltaic Generation	Electricity Net Generation From Wind
2/1/2021	290904.815	9440.997	26672.898
3/1/2021	257708.275	13434.752	39557.527
4/1/2021	240967.586	15516.046	35907.992
5/1/2021	266711.641	17521.734	33191.72
6/1/2021	330638.218	17184.55	26382.741

Data for total renewable and non-renewable production



Representation of renewable and non-renewable energy production

We can see the steady increase in solar and wind when compared to the overall net generation of the data. These are datasets that are used for the project.

4.2. DATA CLEANING

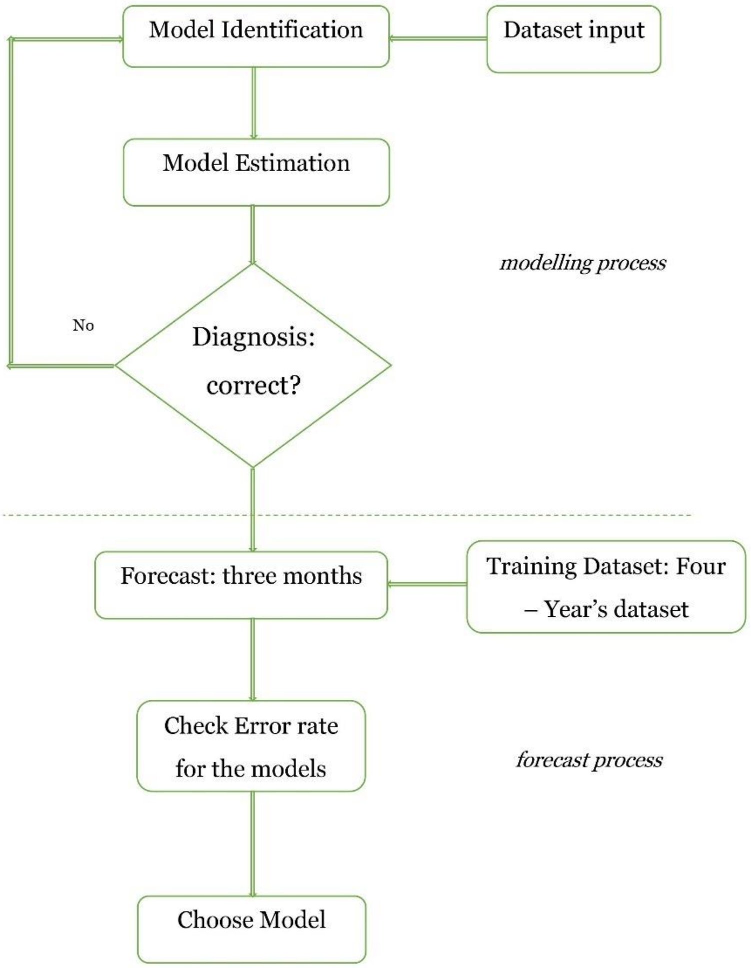
To handle all null values, we can use imputation method for missing value by filling it in using '*fillna()*' function and then drop rows containing missing values. The Column "month" in data set consist of both month and year, we have separated them to different columns - month and year.

The values that are null in columns "Electricity Net Generation Total, Solar Photovoltaic Generation ,Electricity Net Generation from Wind, Total Non-Renewable consumption, Total Renewable Energy Consumption, Solar Energy Consumption, Wind Energy Consumption, Temperature, Solar Photovoltaic Generation are filled using zero. We have considered the dataset from 2010 – 2020 for the consumption of electricity and for production we have data till 2021 June from where we have forecasted the data for the future. We have implemented the algorithms on these datasets after cleaning the data.

5. METHODOLOGY

The methodology used for this project starts with the recognition of data sets needed to start the process and evaluation of data based on the resulting output and expected forecast that the study aims to achieve with the results.

The process stars with the datasets input and collecting the data from various resources based on the availability. These datasets once collected will be identified for further usage and discard the excess data and clean the datasets based on the requirement of project needs.



Flow chart of machine learning algorithm.

Based on the data sets available, we predict the model that suits the datasets for training and testing the datasets. Through the process of trial-and-error methods we choose the possible methods that could be suited to train the dataset on. If the error rate seems to be high, we restart the process of selecting the models to train the algorithm. In this project, the algorithms chosen to train the dataset are timeseries and linear regression.

Once the diagnosis is done, we move further with the algorithms that best suits for forecasting the data. We then check the accuracy of the algorithms that the dataset is trained on in order to select the best method on which the dataset could be further trained and could be applied on real dataset. In this case, timeseries and linear regression is used for the process of choosing the best method for the dataset. Mean absolute error is the means of comparison that is used in this project to select the best model.

5.1. TIMESERIES ALGORITHM

Time series data is a sequence of observation taken from different sequence of time. Numerous sets of data occur as time series. Monthly/ quarterly based data is considered for the project. Timeseries data is used for tracking daily, hourly, or weekly data over seasonality patters. In this project the data used for forecasting the production of electricity for next quarter has seasonality based as time is constant variable this present in the data. The variation of production of electricity using solar power PV is dependent on the weather patters that change based on season. With consideration of dependent variables, the timeseries that is applied for the project is **univariate timeseries**.

Univariate timeseries is a data with a single time dependent variable like demand for a product at time, t. In this case the production of energy and consumption of energy. Both the variables or labels are dependent on time variable. There are many assumptions for this process, other factors like demands will continue to be factor affecting overall product. The univariant time series that is referred, consists of single observations that are noted at same time intervals. This model can usually give within the single number of columns, an implicit variable time is called univariant time series.

Approaches to Univariant time series model.

- Trend, Seasonality, Residual decomposition
- Frequency based Methods
- Autoregressive (AR) model
- Moving Average model
- Box-Jenkins Approach

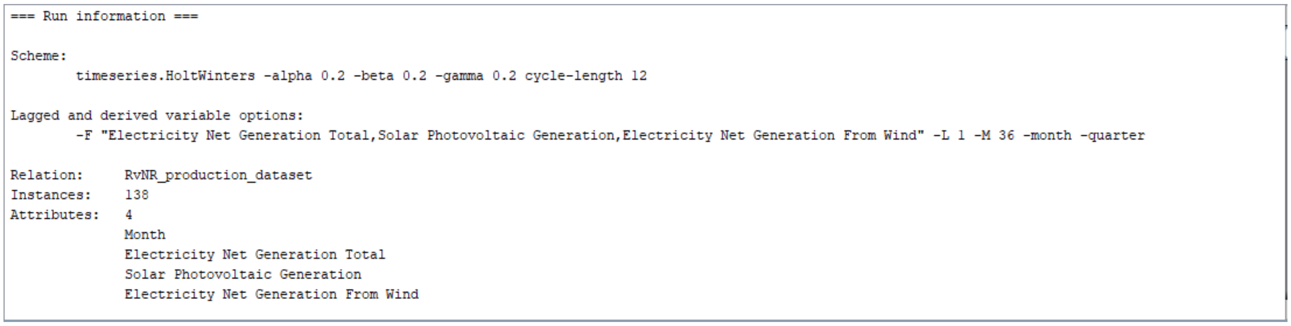
ETS model (Error, trend, seasonality) model is a method that is used for forecasting time series. This model is focused on seasonal components and trend as mentioned by (*Chesilia and others*, *2021*) in '*Selection for the best ETS (error, trend, seasonal) model to forecast weather in the Aceh Besar District*'.

5.1.1. HOLTS-WINTER SEASONAL METHOD

The Holts - Winter method is mainly used to capture seasonality. The process of forecasting the data for production of electricity for both renewable and non-renewable resources is dependent on time variable as previously mentioned. Being the base of forecasting the seasonality that uses three smoothing equations – lt,bt, and st where lt represents level, bt represents trend and st represent seasonality component with corresponding smoothing parameters alpha, beta, and gamma.

The below represents the formula used for holts – winter method

$$\begin{aligned}
\hat{y}_{t+h|t} &= \ell_t + hb_t + s_{t+h-m(k+1)} \\
\ell_t &= \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \\
s_t &= \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m},
\end{aligned}$$



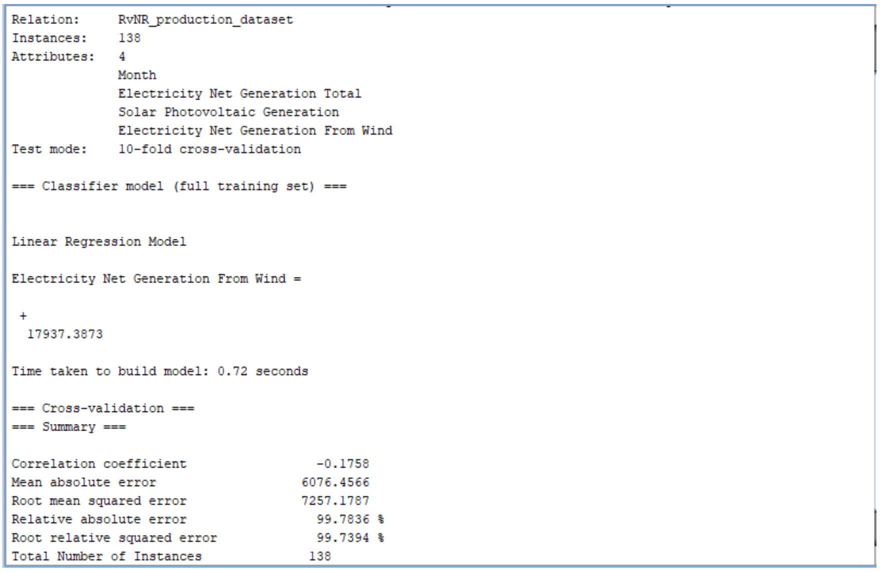
```
=== Run information ===

Scheme:
        timeseries.HoltWinters -alpha 0.2 -beta 0.2 -gamma 0.2 cycle-length 12

Lagged and derived variable options:
        -F "Electricity Net Generation Total,Solar Photovoltaic Generation,Electricity Net Generation From Wind" -L 1 -M 36 -month -quarter

Relation:     RvNR_production_dataset
Instances:    138
Attributes:   4
              Month
              Electricity Net Generation Total
              Solar Photovoltaic Generation
              Electricity Net Generation From Wind
```

The datasets chosen for the project are suitable for applying holts - winter model as it suits the component of forecasting the data further. Among addictive and multiplicative method that are present, additive method is preferred for this process as the seasonality variation is constant for the entire dataset. There is a lag interval that is considered for Holts- winter model that helps in forecasting the data.

5.2.LINEAR REGRESSION

Linear regression is a method for predicting or visualizing the connection between two variables. There are two types of variables considered in linear regression tasks: the dependent variable and the independent variable. The independent variable is a variable that exists independently of the other variables. The levels of the dependent variable will fluctuate when the independent variable is changed. The dependent variable is the variable that is being studied, and it is this variable that the regression model seeks to predict. Every observation/instance in a linear regression job contains both the dependent variable value and the independent variable value.

The equation y= a + bx, where y is the dependent variable (y-axis), x is the independent variable (x-axis), b is the line's slope, and a is the y-intercept. We've performed linear regression on the dataset to predict the Energy consumption for the next few months as mentioned by (*M. Huang,2017*) in the paper "*Theory and Implementation of linear regression"*.



```
Relation:     RvNR_production_dataset
Instances:    138
Attributes:   4
              Month
              Electricity Net Generation Total
              Solar Photovoltaic Generation
              Electricity Net Generation From Wind
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


Linear Regression Model

Electricity Net Generation From Wind =

 +
  17937.3873

Time taken to build model: 0.72 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                 -0.1758
Mean absolute error                   6076.4566
Root mean squared error               7257.1787
Relative absolute error                 99.7836 %
Root relative squared error             99.7394 %
Total Number of Instances              138
```

Data represents the linear regression applied for electricity production

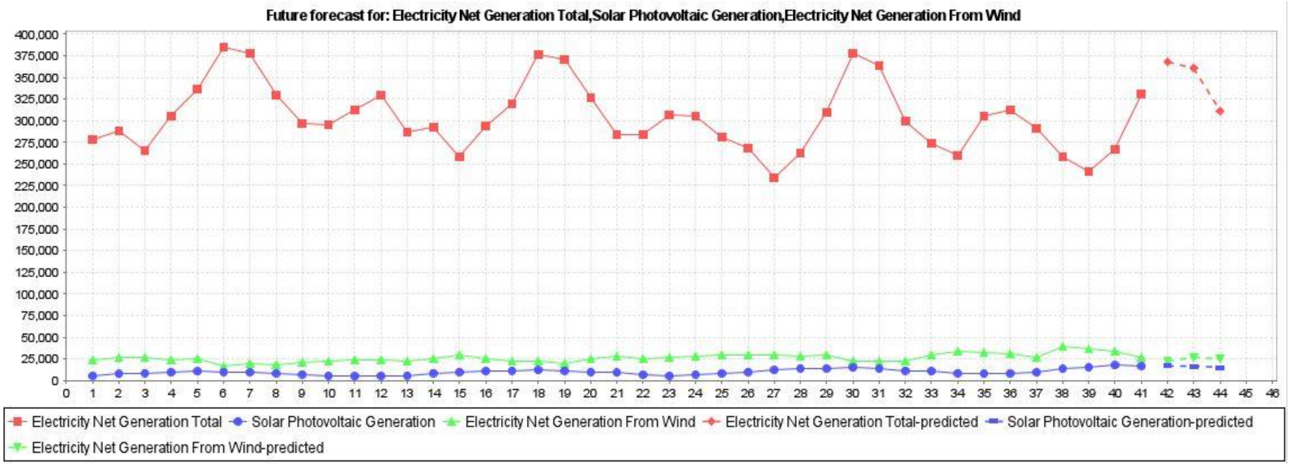
The reason for selecting this algorithm for the project is that it best suits for the dataset after timeseries. With the datasets being dependent on time, the linear regression could be a better algorithm for forecasting the production of electricity and with the comparison of accuracy we get to figure out the better methods to implement the data.

6. EXPERIMENTAL RESULTS

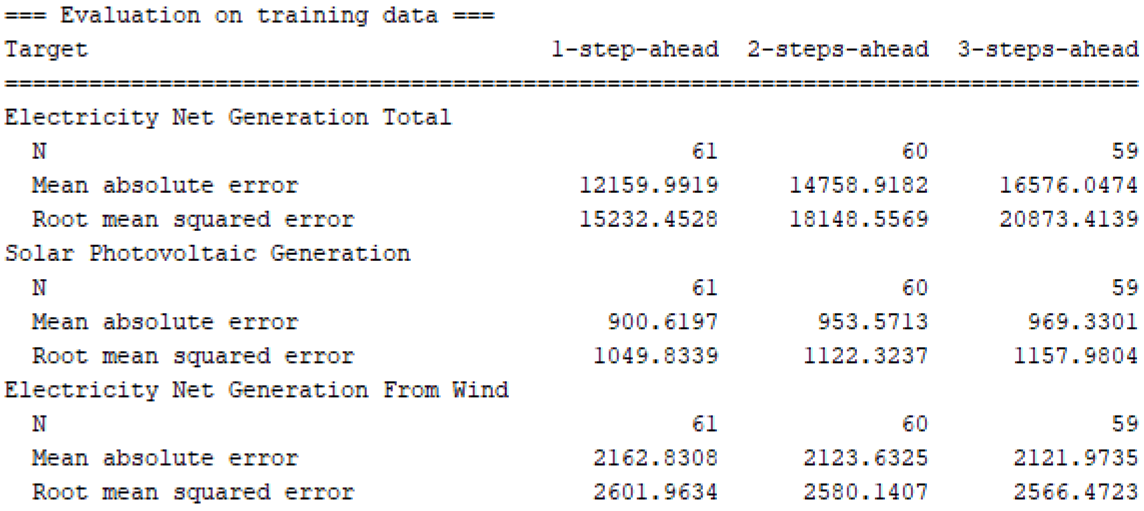
6.1. TIMESERIES

We have used WEKA to train and test the datasets using timeseries and linear regression algorithms. The datasets used in these both cases are mentioned below in the figure. The variables used are Month, electricity net generation, solar electricity generation, wind electricity generation.

Generation data

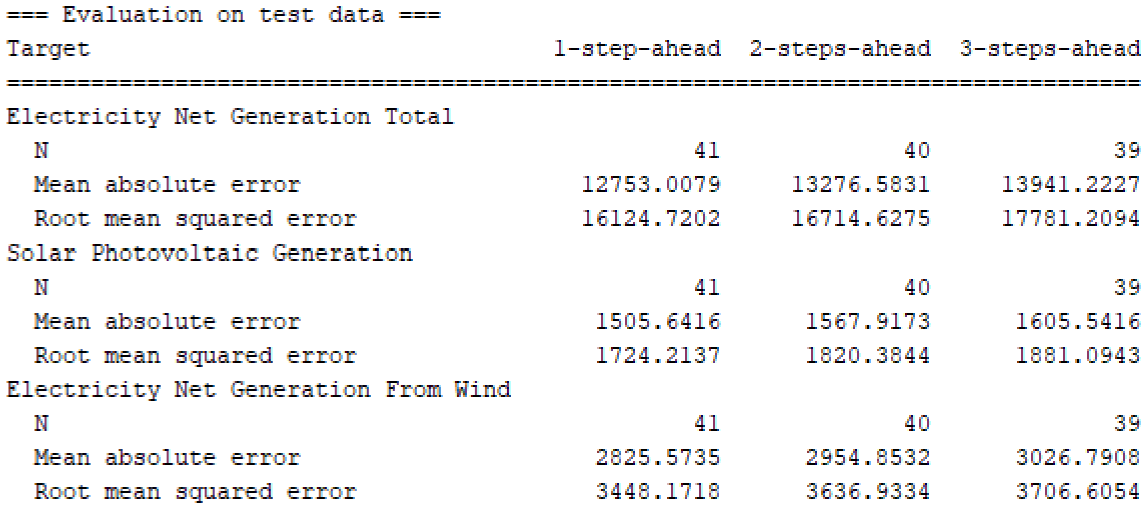


The above data represents the forecast of net generation of electricity using timeseries. We could see that there is decreasing trend for the net generation total showing the actual cause of reducing the coal usage and other non-renewable resources. The data shows that there is a constant trend in solar and wind there is a steady growth representing that there could be further growth.



```
=== Evaluation on training data ===
Target                                1-step-ahead  2-steps-ahead  3-steps-ahead
================================================================================
Electricity Net Generation Total
  N                                             61             60             59
  Mean absolute error                   12159.9919     14758.9182     16576.0474
  Root mean squared error               15232.4528     18148.5569     20873.4139
Solar Photovoltaic Generation
  N                                             61             60             59
  Mean absolute error                     900.6197       953.5713       969.3301
  Root mean squared error                1049.8339      1122.3237      1157.9804
Electricity Net Generation From Wind
  N                                             61             60             59
  Mean absolute error                    2162.8308      2123.6325      2121.9735
  Root mean squared error                2601.9634      2580.1407      2566.4723
```

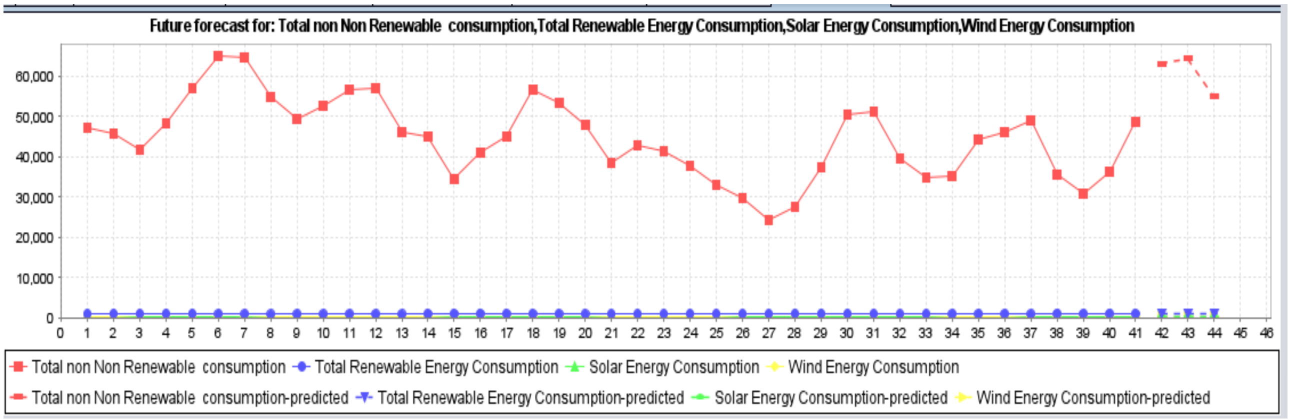
The data above shows the mean absolute error and root mean squared error value of training dataset where the timeseries algorithm has been used to forecast the data for three instances, in this case a total of 97 instances have been used for the study to train and predict the remaining instances.



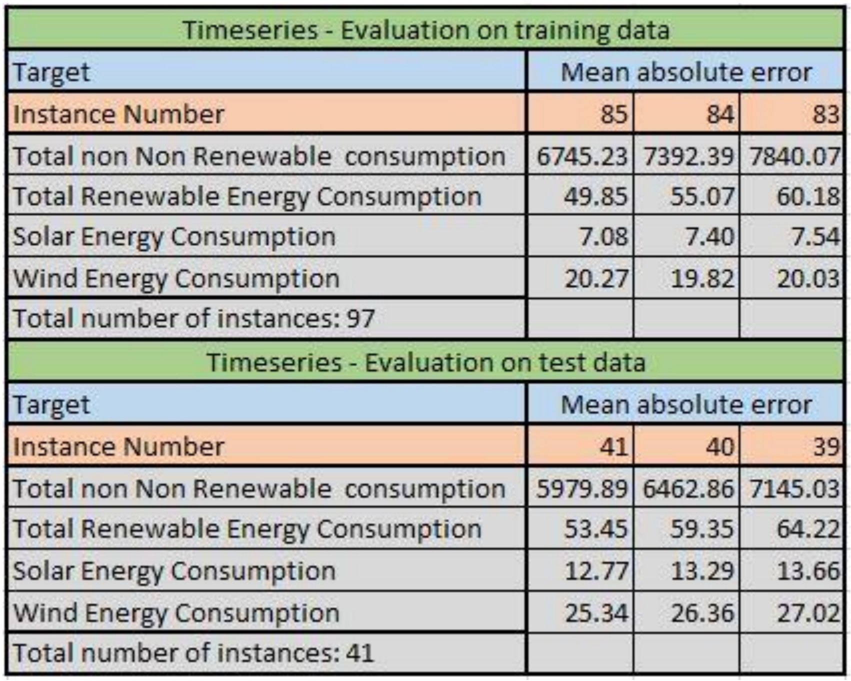
```
=== Evaluation on test data ===
Target                                1-step-ahead  2-steps-ahead  3-steps-ahead
================================================================================
Electricity Net Generation Total
  N                                             41             40             39
  Mean absolute error                   12753.0079     13276.5831     13941.2227
  Root mean squared error               16124.7202     16714.6275     17781.2094
Solar Photovoltaic Generation
  N                                             41             40             39
  Mean absolute error                    1505.6416      1567.9173      1605.5416
  Root mean squared error                1724.2137      1820.3844      1881.0943
Electricity Net Generation From Wind
  N                                             41             40             39
  Mean absolute error                    2825.5735      2954.8532      3026.7908
  Root mean squared error                3448.1718      3636.9334      3706.6054
```

The above data represents the test set along with the overall mean absolute error and root mean squared error value. Based on the two data comparisons, the test set has better results.

Consumption data



The above table represents the overall consumption and comparing the data with other renewable resources. We see that there is decrease in trend for the consumption of overall though there is an initial increase in the data. We also see that there is a constant trend followed for wind and solar.



Timeseries - Evaluation on training data			
Target	Mean absolute error		
Instance Number	85	84	83
Total non Non Renewable consumption	6745.23	7392.39	7840.07
Total Renewable Energy Consumption	49.85	55.07	60.18
Solar Energy Consumption	7.08	7.40	7.54
Wind Energy Consumption	20.27	19.82	20.03
Total number of instances: 97			
Timeseries - Evaluation on test data			
Target	Mean absolute error		
Instance Number	41	40	39
Total non Non Renewable consumption	5979.89	6462.86	7145.03
Total Renewable Energy Consumption	53.45	59.35	64.22
Solar Energy Consumption	12.77	13.29	13.66
Wind Energy Consumption	25.34	26.36	27.02
Total number of instances: 41			

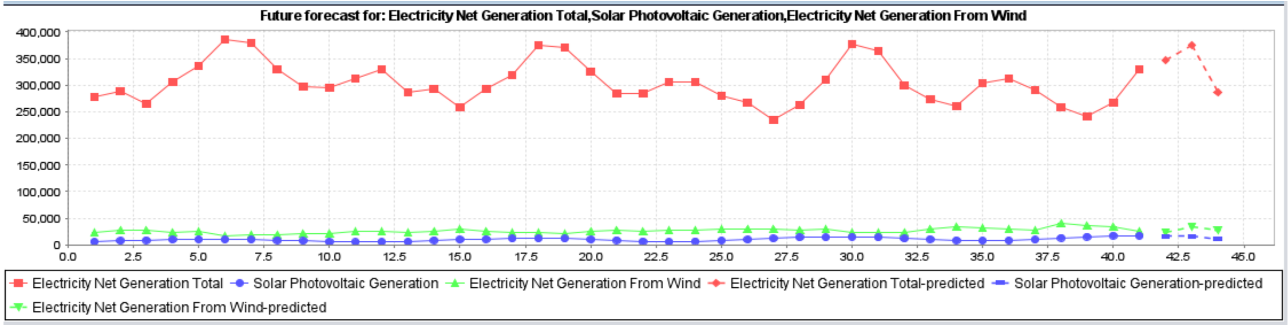
Timeseries data evaluation

From the data above we could see the final evaluation of MAE(Mean absolute error) for timeseries that is performed on training set and test set.

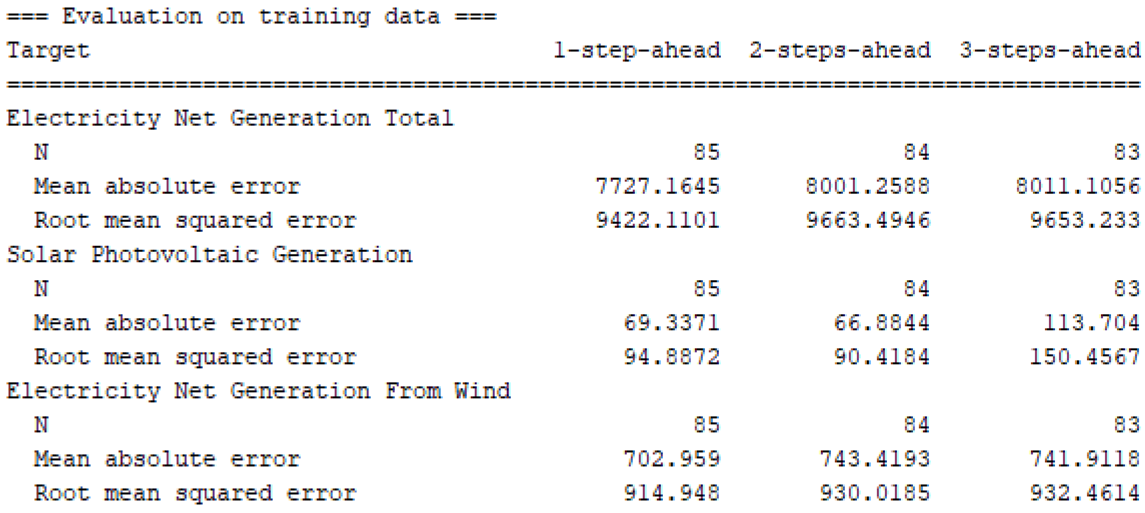
6.2 LINEAR REGRESSION

Linear regression for the project is considered for the comparison purpose to evaluate the basic understating of timeseries as the datasets are clearly dependent on time. With regards to implementation of linear regression, the same datasets that were used for training and test set were used in this case.

Generation Data



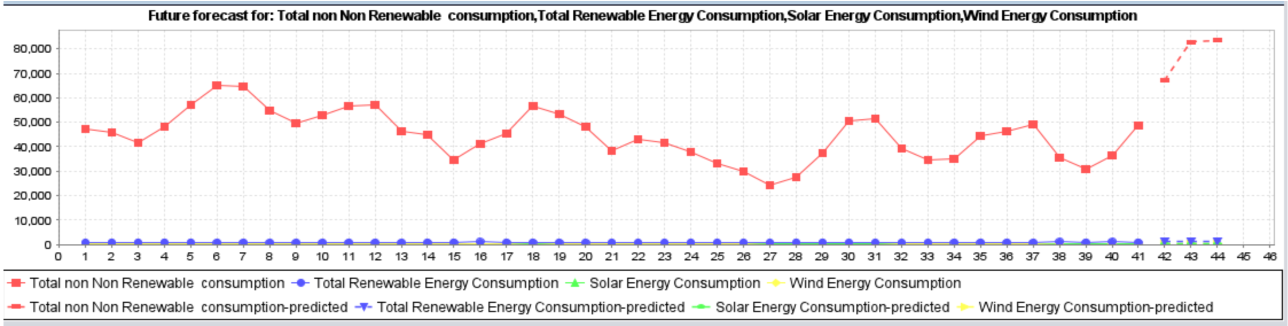
The above graph shows the data of electricity generation and its forecasting done using linear regression, showcasing a similar trend when compared to timeseries. Though they follow similar trend, the accuracy of these two algorithms determines the best algorithm that could be used on real datasets.



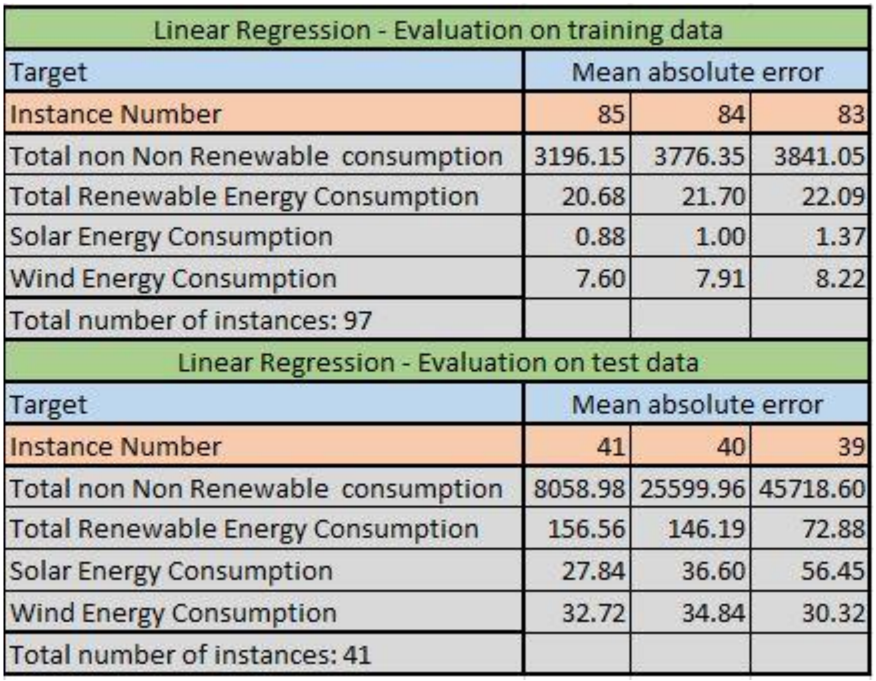
```
=== Evaluation on training data ===
Target                                 1-step-ahead  2-steps-ahead  3-steps-ahead
==================================================================================
Electricity Net Generation Total
  N                                              85             84             83
  Mean absolute error                     7727.1645      8001.2588      8011.1056
  Root mean squared error                 9422.1101      9663.4946       9653.233
Solar Photovoltaic Generation
  N                                              85             84             83
  Mean absolute error                       69.3371        66.8844        113.704
  Root mean squared error                   94.8872        90.4184       150.4567
Electricity Net Generation From Wind
  N                                              85             84             83
  Mean absolute error                       702.959       743.4193       741.9118
  Root mean squared error                   914.948       930.0185       932.4614
```

The data represents the mean absolute error and root mean squared error for the generation of electricity for both renewable and non-renewable resources.

Consumption data

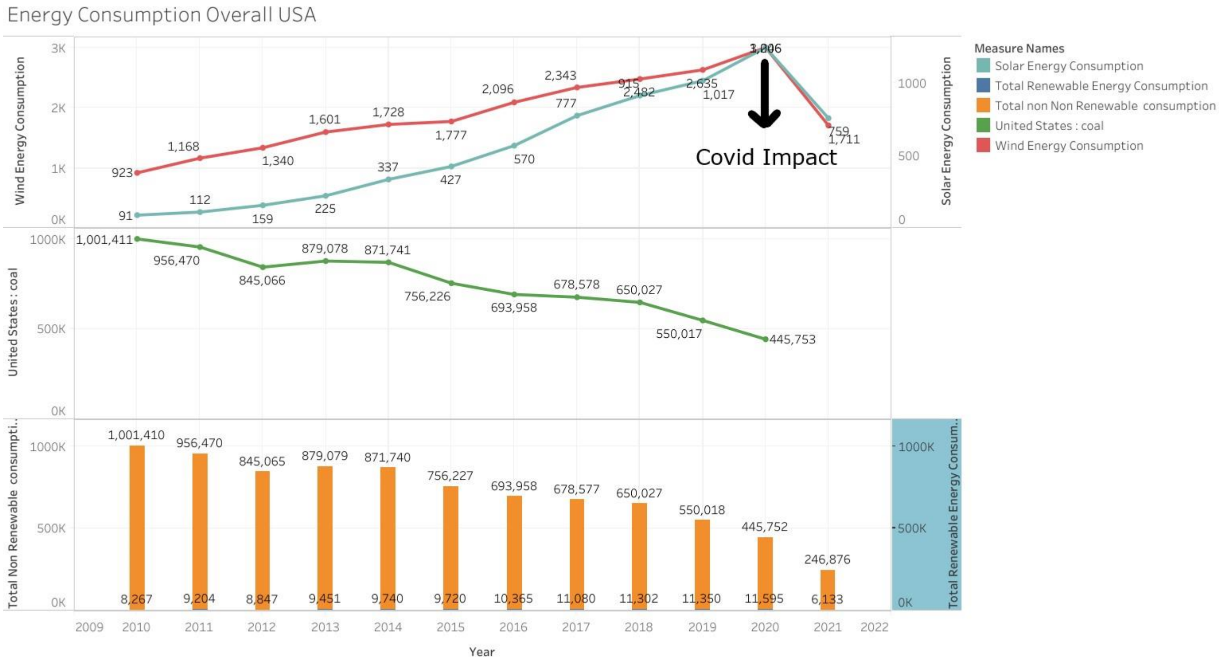


As we could see the forecasting of consumption data using linear regression shows an upward increase for non-renewable resources where there is decrease trend showcased for renewable resources. Compared to timeseries, we see that there is an upward trend indicating that the error rate for the test set in linear regression seems to be high when compared to the error rate obtained for timeseries.



Linear Regression - Evaluation on training data			
Target	Mean absolute error		
Instance Number	85	84	83
Total non Non Renewable consumption	3196.15	3776.35	3841.05
Total Renewable Energy Consumption	20.68	21.70	22.09
Solar Energy Consumption	0.88	1.00	1.37
Wind Energy Consumption	7.60	7.91	8.22
Total number of instances: 97			
Linear Regression - Evaluation on test data			
Target	Mean absolute error		
Instance Number	41	40	39
Total non Non Renewable consumption	8058.98	25599.96	45718.60
Total Renewable Energy Consumption	156.56	146.19	72.88
Solar Energy Consumption	27.84	36.60	56.45
Wind Energy Consumption	32.72	34.84	30.32
Total number of instances: 41			

The data shows the error rate for the linear regression for raining set and test set. We see that the error rate for test set in linear regression seems to be high compared to timeseries indicating that the best model that could be chosen to go further and test on the real dataset is **Timeseries.**



The above data represents the actual electricity consumption in USA and there is a decrease in trend due to the covid impact and indicate that the usage of electricity has been decreased in industries for the past one year. Comparing the future forecast done by both regression and timeseries, the error rate indicates that the **timeseries** showed a better conversion result when the algorithm is applied on test set to forecast next three months when compared with linear regression.

Though the results seem to satisfy the condition of holts-winter timeseries, it is still considered to be **underfitting**. With the datasets being limited to 138 instances, there is still a chance for timeseries algorithm to improve if more instances are considered to train the dataset.

7. CONCLUSION

Timeseries is the most suitable algorithm for this project to implement when compared to linear regression. The project result has provided that the model is still **underfitting** and need to improve further, more datasets need to be considered in order to reduce the **mean absolute error** that is used for the calculation of **accuracy** for **timeseries** and **linear regression**. Based on the final predictions done by timeseries, there is a huge potential for growth of solar power generation and with increase in photovoltaic methods, the future for solar power has tremendous increase in market. Considering the results obtained, it could be said that the **investment in solar energy** compared to other non-renewable energy production sources will be a safe bet for all the investors and potential stakeholders.

8. REFERENCES

1. J. R. Andrade and R. J. Bessa, 2017 "*Improving Renewable Energy Forecasting With a Grid of Numerical Weather Predictions,*" in IEEE Transactions on Sustainable Energy, vol. 8, no. 4, pp. 1571-1580, Oct. 2017, doi: 10.1109/TSTE.2017.2694340.

2. Arrhenius, Svante, 2009, "*On the influence of carbonic acid in the air upon the temperature of the ground*", The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. Retrieved on October September 13, 2021.

3. M. Huang, 2020 "*Theory and Implementation of linear regression,"* 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), 2020, pp. 210-217, doi: 10.1109/CVIDL51233.2020.00-99.

4. IRENA (2021), "*Renewable Power Generation Costs in 2020*", International Renewable Energy Agency, Abu Dhabi. Retrieved on November 20, 2021.

5. Suparna Ray, 2021, "*Renewable's account for most new U.S. electricity generating capacity in 2021*", U.S. Energy Information Administration ,published on January 11, 2021 , retrieved on October 7, 2021

6. UN Environmental programme,2021 "*COP26 ends with agreement but falls short on climate action*" from UNEP, retrieved on November 27, 2021 from *https://www.unep.org/news-and-stories/story/cop26-ends-agreement-falls-short-climate-action*

7. Valéry Masson and others, 2014 *"Solar panels reduce both global warming and urban heat island*", The French National Research Agency for the MUSCADE project, retrieved on September 13, 2021.

8. W. Yan, 2012 "*Toward Automatic Time-Series Forecasting Using Neural Networks*," in IEEE Transactions on Neural Networks and Learning Systems, vol. 23, no. 7, pp. 1028-1039, July 2012, doi: 10.1109/TNNLS.2012.2198074.