Machine Learning for Load Forecasting in Power Systems

Salava V Satyanarayana¹, Pillalamarri Madhavi¹
¹Deparment of Electrical and Electronics Engineering, Hyderabad Institute of Technology and Management
*Corresponding author : pshve2011@gmail.com

Abstract. For the electrical sector, the analysis of massive volumes of data acquired from different electrical systems like Generation, Transmission, and Distribution plays a vital role. Without human interaction, control systems like SCADA and HMI are used to evaluate the data, which is retrieved from various electrical systems such as Generation, Transmission, and Distribution. Automation of every system is necessary to fulfil industry 4.0 criteria. The Internet of Things (IoT) can be used to do this by incorporating the data while implementing proper cybersecurity safeguards. To improve the operational maintenance of electrical systems in the future, this research makes the suggestion that intelligent predictive data analysis be used. Several energy sources and total capacity data files are used in the analysis of both contemporary and historical data in the study. Supervised machine learning algorithms are used to analyze the data that is accessible, and each algorithm's precision is evaluated by the examination of anticipated data.

1 Introduction

Electrical Systems collect massive volumes of data from a variety of sources, including generation, consumption, and transmission [1], in order to minimize human intervention. For many industries, automation of data unification is essential, and technology advancement must take security issues into account. However, connecting every file to one another might make interacting with the cloud more difficult. Systematic data monitoring and control are crucial as the supply and demand of power rise quickly. Typically, the electrical system uses SCADA (Supervisory Control and Data Acquisition) for data analysis. Automation's main objective is to make the electrical system more efficient while preventing data loss, much like a smart system. Future applications, such as Energy Management Systems (EMS) and future electric markets with control algorithms, can benefit from this data analysis. Big data analysis employs a variety of tools and techniques to enable transparency throughout the transferring and linking of the system, which is flexible enough for real-time applications. Feedback information can be used to improve appropriate decision-making algorithms without mistakes or disruptions [2]. More effective results are obtained when regulatory frameworks are integrated with all systems.

In order to perform activities without requiring human interaction, Industry 4.0 demands the integration of artificial devices with physical systems. The economic problem [3], which arises from the requirement to integrate numerous devices with the systems, is one of the main difficulties. Demand-side management, utility consumption, diverse sectors of power generation, and load-side demand prediction are only a few of the key uses of big data analytics [4]. Additionally, it can forecast near-term solar energy and support the management of renewable energy sources for effective energy management.

In recent years, artificial intelligence (AI) has significantly improved the power grid industry. It has made it possible to create electricity grids that are more effective, dependable, and sustainable. Applications of AI algorithms and techniques are being made for generation, transmission, distribution, and consumption in the grid. Demand-side management is one of the main uses for AI in the electrical grid. AI can effectively forecast future energy demands with the aid of machine learning algorithms, allowing grid managers to modify the energy supply accordingly. Balancing the strain on the grid helps to avoid brownouts and blackouts [5]. The optimization of power generation is another way AI is used in the power grid. In order to forecast future energy demand and maximize energy output to fulfill that demand, AI systems can evaluate previous data [6]. As a result, the electricity grid becomes more ecologically friendly and sustainable by reducing its dependency on fossil fuels. AI is also being utilized to increase the grid's effectiveness by spotting and fixing problems in real time. Grid operators can spot potential problems and take corrective action to stop power outages or other disturbances with the aid of AI algorithms.

Overall, incorporating AI into the power grid has the potential to completely transform how we generate, transport, and use energy. It makes it possible for a power grid that is more effective, dependable, and environmentally friendly [7], which is good for consumers and the environment. The integration of renewable energy sources is supported by
the application of AI in the electrical industry, which also increases asset reliability and facilitates efficient energy management. It prepares utilities, operators, and customers with sophisticated systems and decision-making skills to tackle evolving issues in the electrical industry. AI techniques like time series analysis and neural networks are used for load forecasting. These algorithms use historical load data, weather trends, and other relevant factors in order to accurately anticipate future load profiles. Accurate load projections help utilities manage their grid, schedule, and resources more effectively.

2 Machine Learning Algorithms

Any system must be capable of efficiently analysing enormous amounts of data, whether they are structured or not. To accomplish this, it is necessary to include a number of machine-learning approaches in order to categorize and resolve the issue within a set amount of time. Since real-time data analysis [8] frequently includes raw data that is useless to the system, it is difficult and impractical. The data sets can be trained and classified using supervised and unsupervised learning approaches, producing efficient and precise output predictions. Big data analytics is the process of extracting significant information and facts from huge and complex datasets.

Machine learning algorithms are crucial to large data analytics because they can automate analysis, pattern recognition, and prediction tasks, the characteristics of the data, the nature of the problem, and the desired outcomes all have an impact on the selection of an appropriate algorithm. Fraud detection, healthcare and monitoring, energy management, natural language processing in social media, and marketing are applications of machine learning algorithms. These are only a few of the many applications that machine learning algorithms have. As the field matures, machine learning will probably find more applications, advancing technology and decision-making across a range of industries.

Supervised Machine Learning Algorithm: The most popular learning strategy that uses training and testing sets of data is this one. To ensure improved prediction, the input data must first be trained, divided into training and test datasets, and then tested using various techniques. Data can be separated into categories by classification and regression. Regression entails mapping labelled input to continuous output, whereas classification entails mapping labelled input to labelled output. Support Vector Machine (SVM), Decision Tree, Random Forest, and Naive Bayes are the algorithms applied in this method. Regression uses linear and logistic techniques.

- Data gathering: A labeled dataset with each instance connected to the input attributes and output labels for which it is intended is gathered. The dataset consists of two sets: a training set and a test/validation set. Taking important features from the supplied data and extracting them: Depending on how well they predict the target variable, relevant features are either selected or retrieved from the input data. Features can be mixed or altered using feature engineering techniques to improve the model's performance.
- Model training: The model learns the underlying patterns and connections between the input features and output labels using the training set and the selected algorithm. To lower prediction errors, the model's parameters are modified iteratively.
- Model Evaluation: Using the test/validation set, the trained model's performance is assessed. Common evaluation criteria include accuracy, precision, recall, F1 score, and area under the ROC curve (AUC-ROC).
- Model tuning: If the model's performance is subpar, the hyperparameters of the algorithm can be altered to improve performance. Hyperparameters are specified before training to regulate the behavior of the learning algorithm.

After training and evaluation, the model can be used to predict or categorize previously unseen new data by incorporating the patterns found in the data into the input features. It is the basis for a wide range of useful applications, such as spam detection, image recognition, medical diagnosis, and language translation.

Unsupervised Machine Learning Algorithms: In this method, accurate predictions are created by clustering data after unlabeled data sets are analyzed in relation to the prior one. This method is employed to find patterns in data sets that have been incorrectly tagged or classified. A fundamental approach for unsupervised data is K-means clustering. Unsupervised learning techniques are extensively used in exploratory data analysis, data preparation, feature engineering, and data visualization. Although the labels or findings themselves may be confusing, they can nonetheless show underlying patterns and provide useful assessments of the data. Finding innate structures or groups in the data without the use of labels or prior knowledge is the core goal of unsupervised learning. Unsupervised machine learning algorithms are used in many different applications across many different sectors. Several common uses for unsupervised machine learning techniques, such as clustering, video analysis, natural language processing, and data preprocessing.

3 Results and Discussion

The process for projecting power demand trends over time in various states is shown in the diagram above. This involves applying supervised machine learning algorithms to analyse labelled data in order to forecast future trends of electricity generation from different sources in a safe and secure manner [9].
The analysis will shed light on how much energy various states are expected to use in the coming year, and the performance of each algorithm (Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Naive Bayes, and Support Vector Machine) will be compared to determine which one makes the most accurate predictions. This strategy will make the prediction simple using a powerful supervised machine algorithm.

Using Python libraries like Pandas or NumPy to import and deal with datasets in a code cell within the notebook is shown in the aforementioned figures. They are: importing a dataset from a local file (local disk) and importing a dataset through a URL. There are two ways to import datasets utilizing libraries. Libraries like pandas (Provide data structures like Data Frames and Series that make it easier to handle, clean up,
and preprocess structured data) depend on the CSV file (production of electricity from diverse resources). Pandas (Offers support for large matrices and multidimensional arrays, as well as a variety of mathematical operations to efficiently work with these arrays) is widely used to execute operations on datasets such as data cleansing, filtering, grouping, merging, and transformation. NumPy (Offers support for large matrices and multidimensional arrays. NumPy is frequently used for tasks involving linear algebra, numerical computations, and data processing, in addition to seaborn (which provides an intricate user interface for creating engaging and instructive statistical graphics). Seaborn makes it simpler to build sophisticated graphics like heatmaps, violin plots, and pair plots, to name a few.

```python
In [5]: data.isna().sum()
Out[5]:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>EnergySources</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
</tr>
<tr>
<td>dtype: int64</td>
<td></td>
</tr>
</tbody>
</table>

In [6]: data['EnergySources'].value_counts()
Out[6]:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Renewable Energy Sources</td>
<td>4</td>
</tr>
<tr>
<td>Thermal</td>
<td>3</td>
</tr>
<tr>
<td>Thermal Total (A)</td>
<td>1</td>
</tr>
<tr>
<td>Nuclear Total (B)</td>
<td>1</td>
</tr>
<tr>
<td>Nuclear</td>
<td>1</td>
</tr>
<tr>
<td>Renewable Energy Sources Total (C)</td>
<td>1</td>
</tr>
<tr>
<td>Grand Total (A+B+C)</td>
<td>1</td>
</tr>
<tr>
<td>Name: EnergySources, dtype: int64</td>
<td></td>
</tr>
</tbody>
</table>
```

**Fig. 3** Retrieving the Data from the Data Set

```python
In [8]: sns.countplot(data['EnergySources'], label='count')
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x2a7beab2c08>
```

**Fig. 4** Graphical Representation of Data Using Matplotlib

```python
In [9]: # Split the data into independent(x) and dependent(y) data sets
   x = data.drop(['EnergySources'], axis=1)
   y = data.EnergySources.values

In [10]: # Split the data set into training 75% and testing 25% data
   from sklearn.model_selection import train_test_split
   x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.25, random_state=0)
```
Fig. 5 Testing and Training the Data

```python
In [11]:
print("x_train: ", x_train.shape)
print("y_train: ", y_train.shape)
print("x_test: ", x_test.shape)
print("y_test: ", y_test.shape)

x_train: (9, 1)
y_train: (9,)
x_test: (3, 1)
y_test: (3,)
```

Fig. 6 Testing the Data for Different Supervised Machine Learning Algorithms

```python
In [12]: def models(x_train, y_train):
    
    # Logistic regression
    from sklearn.linear_model import LogisticRegression
    log = LogisticRegression(random_state=0)
    log.fit(x_train, y_train)

    # Decision Tree
    from sklearn.tree import DecisionTreeClassifier
    tree = DecisionTreeClassifier(criterion='entropy', random_state=0)
    tree.fit(x_train, y_train)

    # Random Forest Classifier
    from sklearn.ensemble import RandomForestClassifier
    forests = RandomForestClassifier(n_estimators=10, criterion='entropy', random_state=0)
    forests.fit(x_train, y_train)

    # Gaussian Naive Bayes
    from sklearn.naive_bayes import GaussianNB
    nb = GaussianNB()
    nb.fit(x_train, y_train)

    # SVM
    from sklearn.svm import SVC
    svm_clf = SVC()
    svm_clf.fit(x_train, y_train)

    #print the model accuracy on the training data
    print('[0] Logistic Regression Training Accuracy: ', log.score(x_train, y_train))
    print('[1] Decision Tree Classifier Training Accuracy: ', tree.score(x_train, y_train))
    print('[2] Random Forest Classifier Training Accuracy: ', forests.score(x_train, y_train))
    print('[3] Naive Bayes Training Accuracy: ', nb.score(x_train, y_train))
    print('[4] SVM Training Accuracy: ', svm_clf.score(x_train, y_train))

    return log, tree, forests, nb, svm_clf
```

Fig. 7 Accuracy for Different Supervised Machine Learning Algorithms

```python
In [13]: models(x_train, y_train)

[0] Logistic Regression Training Accuracy: 0.3333333333333333
[1] Decision Tree Classifier Training Accuracy: 1.0
[2] Random Forest Classifier Training Accuracy: 0.8888888888888888
[3] Naive Bayes Training Accuracy: 0.7777777777777778
[4] SVM Training Accuracy: 0.4444444444444444
```
The above figures clearly represent the accuracy of different supervised Machine Learning Algorithms [10] to predict the power from generating sources like Hydro, Thermal-Steam, Thermal-Diesel, Thermal-gas, Nuclear, Renewable, and overall capacity for future purposes all over India. The objective is to predict future data associated with the "Source-wise installed generation capacities in the country as of 30.06.2022" data file, by utilizing machine learning algorithms such as Logistic Regression, which calculates the probability for one to all by combining all decisions, giving an accuracy of 0.333. Using a data collection and precise decisions made under probabilistic settings, the Decision Tree Classifier calculates accuracy as 1.0. Combining all decisions that provide an accuracy, the Random Forest Classifier determines the chance for one to all with 0.888, Nave Bayes Classifier, which employs the Bayes Theorem to determine the likelihood of each class to produce the primary class and provides an accuracy of 0.777, Support Vector Machine, which uses hyperplanes to create functions and provides an accuracy of 0.444. By implementing all the algorithms, the Decision Tree Classifier provides better efficiency and higher accuracy.

4 Conclusion

Based on a particular dataset (Power generation from various sources), predictions can be produced for upcoming years using machine learning algorithms. In the case of the aforementioned dataset, multiple algorithms produced varying degrees of accuracy, indicating Decision Tree Classifier and Random Forest Classifier can be used to predict power generation from various energy sources for the following years. Therefore, it can be inferred that applying machine learning algorithms to power systems reduces the complexity of data processing and improves the precision of predictions.

5 Future Scope

Machine learning's future in load forecasting for power systems seems positive, with potential for innovation, better accuracy, and increased reliability. It is critical to handle the power industry's developing needs, such as the integration of renewable energy sources and the growing need for real-time and precise load forecasts for efficient and sustainable operation and also ensuring the security of load forecasting algorithms and the data they use. Future research could concentrate on making these models more resistant to hackers.

6 References