Chapter 3

Design and Development of Model Using Data Mining

3.1. Introduction to Modelling
3.2. Types of Models
3.3. General Approach to Build a Model
3.4. Research Methodologies
3.5. Proposed Model for Power Consumption Prediction
   3.5.1 Proposed Model for Agriculture Power Consumption Prediction
   3.5.2 Proposed Model for Domestic Power Consumption Prediction
3.6. Summary
Chapter 3

Design and Development of Model Using Data Mining

3.1 Introduction to Modelling

To build a statistical model of future behaviour through predictive analysis, predictive modelling can be helpful as a process. The prominent area of data mining is to forecast probabilities, trends, and predictive analysis.

In general, a model that focuses on prediction is known as a ‘predictive model.’ This kind of model is made up of the number of predictors, which are variable factors that are likely to impact future behaviour or results.

Historical data is essential in model building. The data is gathered for the appropriate predictors and on the basis of this statistical model is formulated. From the statistical model, predictions are on hand and then the model is either authenticated or reforms for further data to become available.

3.2 Types of Models

In data mining, there are mainly three kinds of models, i.e., predictive model, description mode, and decisive model. All these types are discussed below:

3.2.1 Predictive Model

The key goal of this model is to obtain the likeness in performance in one sample present in another sample. To seek out subtle data patterns or to find hidden patterns or relationships between data, models have their applicability in many areas.

The regression model is a statistical procedure, which allows a researcher to estimate the linear or straight line, a relationship that relates two or more variables. To summarize, the huge amount of change in one variable that is associated with a change in another variable or variables, a linear relationship can be the key feature for the researcher.

When your main goal is a prediction, cross-validation help to select a predictive model and also to check how accurately that model will play its role in training data set. In prediction, a model will use two kinds of data set, which is known data and unknown data commonly referred to as training data set and test data set respectively.

3.2.2 Descriptive Model

By using descriptive model, compute relationships in data can be classified. The main difference between descriptive model and the predictive model is that the descriptive model focuses on predicting many different relationships identified by this model.

Predictive models arrange data by their possibility of taking a particular action the way descriptive models do not. Instead, descriptive models can be used to classify data by their first choice and life stage. These kinds of models are used to cluster large data items into groups and then each group is further fed up in the predictive model.
3.2.3 Decisive Model

Decision model in which relationships among all the elements of a decision are described. In this model, many variables require training data, decision and forecast results to predict the result. These kinds of models are used to optimize decision logic or a set of business rules that will generate the desired action for every situation.

3.3 General Approach to Build a Model

A regression technique is a logical approach to build a regression model basis on a given dataset. Model induction techniques like neural networks, support vector machines, etc. are based on regression. Each technique makes use of a learning algorithm to recognize the model. The model generated by a learning algorithm should match the input data and perfectly predict the previously unknown output values. The process has been shown in the figure below.

The main objective of the learning algorithm is to build a model that have high-quality performance and generalization and regression capabilities. So conclusively, a model must predict the values of previously unknown records.

The future model must be trained against the historical data of the same domain, to get the high-level accuracy in prediction of values.

This training phase finds out the correctness of the model to predict categorical class labels for the unidentified records. After the suitable training of the model, the model can be tested for the unidentified samples.

In order to flag problems like selection bias or overfitting, cross-validation checks predicted data against the model to be tested on the dataset and give a close look at how the model will simplify to an independent dataset. Cross-validation groups dealings with capability in prediction to obtain a more accurate estimate of model prediction performance.
Figure 3.1 - Categories of Data Mining Algorithms
### 3.3.1 Regression-A Modelling Technique

Generally, regression in data mining used to predict the values. Regression involves predictor variable (the values which are known) and the response variable (values to be predicted). In more general sense, we can say that regression is data mining function, which is used to predict the numbers, i.e., profit loss, product price, sales, house values, temperature, weather, electricity demand using regression technique. There are many other factors where the regression model could be used to predict the values.

To build data, a regression algorithm guesses the value of the target as a function of the predictors for each case. Predictors and target function makes one model and then applied to a different data set in which the target values are unidentified. There are various families of regression functions and different ways of measuring the error.

**Linear Regression** is a straightforward form of regression. It attempts to model the connection between two variables by fitting a linear equation to observed data. On the off chance if the result is a straight line, then it is considered as a linear model and on the off chance that it is a curved line, at that point it is a non-linear model.

If there is no linear dependence between data, it is called as **Non-Linear Regression**. The equation that does not follow the rule of a linear model called a non-linear regression model. **Multivariate Regression** is an extension of linear regression analysis. To predict an outcome and a single continuous dependent variable Multivariate regression model uses two or more independent variables.

Regression analysis is one of the most popular statistical techniques used for predictive modelling and data mining task. Usually, linear and logistic regressions are the first priority algorithms when learning predictive modelling. There are many forms of regression which can be performed to predict forecasting results. Each of them has its own importance and a specific condition where they can be the best suit to apply. Regression analysis is also used to understand the forms of relationships among the variables are related to each other.

Therefore, when the set of samples are to be scattered under earlier known criteria and label then, it is called supervised learning.

### 3.3.2 Regression Model Performance Improvement:

To improve the performance of mode, we used cross-validation technique. Older technique split the data set in percentage for training data set and testing data set. This new cross-validation technique improves performance. When the performance is improved, it reflects in other evaluation matrixes like the correlation of attributes become strong and various errors are reduced. The cross-validation technique and evaluation matrixes are discussed in the following section.

#### 3.3.2.1 Ten fold Cross-validation Technique

Cross-validation is a statistical method used to approximate the skill of machine learning models. Cross-validation is commonly used in applied machine learning to match and select a model. For predictive modelling problem, cross-validation can be helpful because it is easy to understand, easy to implement and result in skill estimates that generally have a lower bias than other methods.

Cross-validation is used to evaluate machine learning models on a limited dataset. The technique is also called rotation estimation or out of sample testing. This technique is any of various related model validation techniques for assessing how the result of a statistical study
will simplify to an independent dataset.

When your main goal is a prediction, cross-validation helps to select a predictive model and also in checking how accurately that model will perform in practice. In prediction, a model is usually given a dataset of known data (called as training data set) and dataset of unknown data (called as test data set) on which model is to be tested.

In ten-fold cross-validation, the original sample is randomly partitioned into ten equal size sub samples, it means to break data into sets of size n/10. From the ten subsamples, a single subsample is kept as the validation data for testing the model, and the left behind 9 (10-1) subsamples are used as training data that test on a single subsample. The cross-validation process is repeated ten times (the folds), with each of the subsamples used exactly once as the validation data. The result from the ten folds can then be averaged or combined to produce a single estimation. The main benefit of this model is that all records are used for both training and validation, and each of the records is used for validation data only one time.

In order to flag problems like selection bias or overfitting, cross-validation checks predicted data against the model to be tested on the dataset and gives a close look at how the model will simplify to an independent dataset — cross-validation groups’ measures of fitness in prediction to obtain a more exact estimate of model prediction performance.

### 3.3.2.2 Evaluation Matrixes

Evaluation matrixes are used for evaluation of the model’s result. Using these various matrixes, one can determine the best performance of algorithms. In these matrixes, one matrix shows the correlation among the attributes and other are shows the error rate in the model. Various evaluation matrixes are as follows.

#### 3.3.2.2.1 Correlation Coefficient

A correlation coefficient is a measurement of the relationship between variables. It shows the strength of the relationship, in which it focuses on relative movement of the two variables. There are two types of correlation coefficient – positive and negative. The boundary of the correlation coefficient is -1.0 to 1.0. If the cost of the correlation coefficient is out of this boundary, then this correlation measurement is incorrect. The formula of correlation (R) is shown below:

$$ R = \frac{\sum dydx}{\sqrt{(\sum dx^2 \sum dy^2)}} $$

#### 3.3.2.2.2 Mean Absolute Error

The mean absolute error value is the arithmetic mean of an error value. This is a measurement of the average magnitude of the error in prediction. In other words, this is the absolute variation between actual and forecasted value. If the value of error is near to zero (0), it indicates prediction is better. The equation used to find out mean absolute error is shown below:

$$ MAE = \frac{1}{n} \sum_{j=1}^{n} |y_j - \hat{y}_j| $$

#### 3.3.2.2.3 Root Mean Squared Error

Root mean squared error is the standard difference of the prediction errors. This is a rule of average magnitude error measurement. It is a square root of the average of squared difference of actual and predicted value. In general, Root mean squared error measure how data points
are far from the regression line. This error is used for verification of the experimental result. The formula of root mean squared error is as follows.

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}
\]

### 3.3.2.4 Relative Absolute Error
Relative absolute error is the same as a relative squared error. This error is related to a simple prediction. Simple prediction is just the average of actual values. Absolute error is the magnitude difference of actual value and predicted value. The relative error is the absolute error divided by the magnitude of actual value. Following equation is used to find out the relative absolute error.

\[
RAE = \frac{\sum_{i=1}^{d} |y_i - y_i'|}{\sum_{i=1}^{d} |y_i - \bar{y}|}
\]

### 3.3.2.5 Root Relative Squared Error
Root relative squared error is a square root of relative squared error. Relative squared error is a total squared error and it is normalized by dividing by the total squared error of simple predictor. The formula of root relative squared error is shown below:

\[
RRSE = \sqrt{\frac{\sum_{i=1}^{d} (y_i - y_i')^2}{\sum_{i=1}^{d} (y_i - \bar{y})^2}}
\]

### 3.3.2.6 Percentage Error
The percentage error is the variance between the actual and predicted value in percentage. This error allows you to see how far off is your predicted value. Percentage error is always expressed as a positive number. The purpose of percentage error calculation is to know how close predicted value is to a true actual value. The formula of percentage error is shown below:

\[
PE = \left(\frac{Y_t - F_t}{Y_t}\right) \times 100
\]

### 3.3.2.7 Mean Absolute Percentage Error
The mean absolute percentage error shows the error in percentage form. It is an average of the percentage error. Sometimes mean absolute percentage error is known as mean absolute percentage deviation. This error is used to see the accuracy of the prediction. In data mining and machine learning fields, it is considered as a loss function for regression. In short, the mean absolute percentage error is used for model evaluation. It is a very instinctive explanation in terms of relative error. The formula of mean absolute percentage error is as follows.

\[
MAPE = \frac{1}{n} \sum_{n=1}^{n} |PE|
\]
3.4 Research Methodologies

A research methodology is a tool in research, which directs the activities in the proper direction, to get research objective. It is also a linked activity from the beginning to the end to achieve the research objective. With the help of specific methods, the researcher formulates the problem in the study. The proposed research is carried out with the following phases:

3.4.1 Quantitative Research Methodology: This method focuses on objective measurements and numerical analysis of data. The proposed research is the foundation of experimental research. The proposed research analyzes several techniques associated with Data Mining techniques and performs quantitative experiments on suitable technique.

3.4.2 Survey Research Methodology: The proposed research performs many reviews like the literature review and data collection survey to reach a specific conclusion. It comprises a literature survey for the use of Data Mining, techniques of Data Mining, different algorithms used in prediction technique and Data collection survey for secondary data.

3.4.3 Formal Methodology: The proposed research work uses already available algorithms and provides an explanation to them in terms of a literary survey or quantifiable approach.

3.4.4 Experimental or Simulation Methodology: The proposed research work estimates new solutions for problems based on experimental methodology.

3.5 Proposed model for power consumption prediction
The following section shows the two proposed models for power consumption prediction one for agriculture power consumption prediction and the second one for domestic power consumption prediction.
3.5.1 Proposed Model for Agriculture Power Consumption Prediction

The following figure shows the proposed model for power consumption prediction for agriculture land.

![Proposed Model for Agriculture Power Consumption Prediction](image)

**Figure-3.2 Proposed Model for Agriculture Power Consumption Prediction**

3.5.1.1 Data Source

Generally, in the simplest term data source is a source of the data. It can be a file, a specific database on a DBMS, historic data of institute, or even a live data. The data might be stored on a computer as a flat file, excel file, XML file, web services or a database file, etc.

Data is one of the most important and essential aspects of any research studies. Therefore, the researchers must choose the proper data source to collect the data. For the research, there will be several data source available. Here, past data of PGVCL and past data of weather are used as data source1 and data source2 respectively in our proposed model for agriculture power consumption prediction.
3.5.1.2 Data Extraction, Transformation, Cleaning and Arranging
Data are the backbone of any research studies. Concerning data, there are many things that can go wrong; it can be construction, arrangement, formatting, spellings, duplication, extra spaces, and so on.

For getting good results from the applied model in data mining and machine learning projects the arrangement of the data must be in the correct way. Various data mining and machine learning algorithm requires information in a particular format. For example, the Random Forest algorithm does not work on null values. Thus, if we want to execute random forest algorithm null values has to be handled from the initial raw data set.

Data pre-processing is required because of the existence of unformatted real data. Most of the real data is collected with inaccurate data or we can say missing data, the existence of noisy data, unpredictable data. Therefore, in our proposed model for agriculture power consumption prediction, data must be pre-processed and clean before final use.

3.5.1.3 Handling Missing Values
In real-world data, there are some instances where a particular record or set of records are absent because of many reasons, such as corrupt data, failure to load the information, or incomplete extraction. Hence, handling the missing values is one of the challenging tasks for researchers. There are various techniques to handle missed data, for example, deleting rows, replacing with mean/median/mode, assigning a unique category, predicting the missing values, using algorithms that support missing values. Here, the researchers used mean value principle technique to replace missing records in our proposed model for agriculture power consumption prediction.

3.5.1.4 Data Normalization
Data normalization is the process of data arrangement dealing with its relations to discard the recurrence of data which are used to build a relationship between entities. Normalization is helpful for the prediction or forecasting use. To sustain the large variety of prediction and forecasting the normalization technique is mandatory to make them faster in order to get a good result. There are some normalization techniques available like min-max normalization, z-score normalization, and normalization by decimal scaling. In our proposed model for agriculture power consumption prediction, we apply the filter facility of researcher tool. Research tool has the facility of filter, including normalization and standardization. Normalization converts the value of all attributes between 0 and 1.

3.5.1.5 Run/Evaluate Prediction Models
In predictive modelling, performance evaluation act as the main role. The implementation of a predictive model is calculated and judged by choosing the right metrics. The choice of metrics influences how the implementation of a predictive model is measured and judged. Without choosing the correct metrics that measure how accurate the model is predicting our problem, we might build a robust model. Proper predictive models evaluation is also important because we want our model to have the same predictive ability across many different data sets. The predictive model can solve two kinds of problem regression and classification. In our predictive model, we run on 16 different regression algorithms each with ten-fold cross-validation technique.
3.5.1.6 Best Model Selection

The model selection is the process of choosing a model between different data mining and machine learning approaches. To find the best model, various evaluation metrics used such as RMSE, RRSE, MAE, RAE, Percentage error, and MAPE. Based on these matrixes, finally selected two regression algorithms (SMOreg, Linear Regression) to apply on real data and choose the best fit model for agriculture power consumption prediction.

3.5.1.7 Testing of models on real data

The real data is used to assess how well the algorithm was trained, and to estimate model properties. In this research, real data is tested on two regression algorithms: SMOreg and Linear Regression.

3.5.1.8 Final Model Generation

To come up with the final model, Accuracy of two selected test model has been considered. After comparing the accuracy of models, choose the best one from that and generate the final model for agriculture power consumption prediction.
### Proposed Model for Domestic Power Consumption Prediction

The following figure shows the proposed model for power consumption prediction for domestic use.

![Proposed Model for Domestic Power Consumption Prediction](image)

#### 3.5.2.1 Data source

Generally, the data source can be a file, a database file on a DBMS, historic data of institute, or even a live data. The data might be stored on a computer as a flat file, excel file, XML file, web services or a database file.

Data is one of the most critical and essential aspects of any research studies. The researchers must choose the proper data source to collects the data. For the research, there will be a number of data source available. Here, the historic data of PGVCL, historic data of weather, population survey and electronic devices are used as data source1, data source2, and data source3 and data source4 respectively in our proposed model for domestic power consumption prediction.
3.5.2.2 Data Extraction, Transformation, Cleaning and Arranging

Data are the backbone of any research studies. Concerning data, there are many things that can go wrong; it can be construction, arrangement, formatting, spellings, duplication, extra spaces, and so on. To achieve better results from the applied model in Machine Learning algorithm, the arrangement of the data is to be in an appropriate manner. Many Machine Learning models should run with proper information, for example, Random Forest algorithm does not work with null values, therefore to apply random forest algorithm null values must be discarded from the original raw data set.

Pre-processing is necessary to handle unformatted real-world data. Most of the real-world data comes up with an incorrect data or we can say missing data, the presence of noisy data, inconsistent data. Hence, in our proposed model for domestic power consumption prediction data must be pre-processed and clean before final use.

3.5.2.3 Handling Missing Values

In real-world data, there are some instances where a particular record or set of records are absent because of many reasons, such as corrupt data, failure to load the information, or incomplete extraction. As a result, handling the missing values is one of the challenging tasks for researchers. There are various techniques to process missed data, for example, Deleting Rows, Replacing With Mean/Median/Mode, Assigning An Unique Category, Predicting The Missing Values, Using Algorithms Which Support Missing Values. Here, the researchers have used Mean Value Principle technique to replace missing records in our proposed model for domestic power consumption prediction.

3.5.2.4 Data Normalization

Normalization is required to avoid the repetition of data. It can be useful for the prediction or forecasting purpose. Before applying prediction and forecasting, the normalization technique is required to make them nearer in order to achieve a better result. Many normalization techniques are available like min-max normalization, z-score normalization, and normalization by decimal scaling in our proposed model for domestic power consumption prediction; we have applied the filter facility of researcher tool. Research tool has the facility of filter, including normalization and standardization. Normalization converts the value of all attributes between 0 and 1.

3.5.2.5 Run/Evaluate Prediction Models

To forecast future behaviour, predictive modelling plays an important role. The completion of a predictive model is calculated and judged by choosing the accurate metrics. Choice of metrics influences how the performance of the predictive model is measured and compared. Without choosing accurate metrics that measures how accurate the model is predicting our problem, we might build a robust model. Proper predictive models evaluation is also necessary because we want our model to have the same predictive ability across many different data sets. A predictive model can solve classification and regression problems. The predictive model runs on 16 different regression algorithms each with 10 fold cross-validation technique.

3.5.2.6 Best Model Selection

Model selection is the process of choosing a model between different machine learning approaches. To find the best model, various evaluation metrics used such as RMSE, RRSE, MAE, and RAE. Finally, we selected three regression algorithms (Random Forest, SMOreg,
and M5P) to apply on real data and select the best fit model for domestic power consumption prediction.

3.5.2.7 Testing of Models on Real Data

The real data is used to assess how well the algorithm was trained, and to estimate model properties. In this research, real data has been tested on three regression algorithms: Random Forest SMOreg and M5P.

3.5.2.8 Final Model Generation

To come up with the final model, accuracy of three selected test model has been considered. After comparing the accuracy of models, we have chosen the best one from the set and generated the final model for domestic power consumption prediction.

3.6 Summary

This chapter deals with the design and development of a predictive model and final model generation based on regression algorithms. It also discusses types of models, regression as a modelling technique in which various algorithms had been run to find the best fit model. For the selection of the best fitted model, the evaluation and performance of matrixes are required, so in this chapter, we have also discussed various performance matrixes with their formulas. Finally, we have discussed two proposed model – a proposed model for agriculture power consumption prediction and proposed model for domestic power consumption prediction including detail description of all steps of models.