Welcome to the presentation

My name is Abdur Rahman, Research Assistant at Energy Informatic Group, Lahore University of Management Sciences.

Along with my co-authors Ameera Arif and Ahmed Nadeem, my fellow research associates at LUMS and our supervisor Dr. Naveed Arshad, I would like to present to you the topic: Modelling of Residential Scale consumer demographics using monthly electricity data.

**[slide]**

First and foremost, I would like to tell you about our motivation for choosing this topic.

Energy in the form of electricity is an ever-increasing requirement. The average consumption of electricity has almost doubled since 2000. While most of the developed countries have been able to handle this increasing demand quite well, developing countries such as Pakistan have faced several problems coping up with it.

These problems include but are not limited to: electricity shortfall on regular basis, overloading of distribution networks, power theft and regular load shedding in residential supply of electricity especially in summer seasons. Moreover, most of the efforts directed to analyze these issues have been directly impeded by the lack of available data.

While the application of advanced infrastructure such as smart grids and smart meters can potentially solve many of the problems, these solutions require massive budget which the already overburdened economy cannot afford.

**[slide]**

Therefore, we present a novel technique to address many of these problems with the existing available resources. Most of the work focused in this area has been limited to prediction and forecasting of consumption patterns given historic data. Our work is a lot different from that.

Our work focuses on prediction of consumer demographics from monthly electricity consumption data. This can potentially make significant improvements to existing load forecasting models. On policy level, the knowledge of consumer demographics can help in resolving policy failures and implementation gaps, as well as improve further development plans and help in solving issues such as power theft.

**[slide]**

So the first step in our methodology was the exploratory data analysis to visualize the data and to check its integrity. The dataset contained some missing information so an intermediate step of data cleaning was required. In this step we removed the houses which had missing electricity consumption data points. Then the cleaned dataset is split into training data and test data for the machine learning models.

To model the relationship between consumer demographics and consumption patterns, we applied three machine learning techniques: Multivariate linear regression, Support vector regression and Neural networks. The dataset we had was split into training and test datasets. The models were trained through the training dataset and their performance was evaluated through the test data.

Let us dive further into handling of our dataset.

**[slide]**

In our work, we have used PRECON dataset holds significance for our work due to a large number of diverse households included as well as the duration of one complete year. This dataset has been collected from Pakistan so it is able to adequately represent the dynamics of a developing country as well. The dataset includes, electricity consumption of 42 houses over a period of one year. So all the seasonal electricity consumption pattern is present in the dataset. The dataset also includes 26 metadata parameters for all the houses. These include the number of air conditioners, the number of fans and many more.

In the figure on the right, a box plot shows the diversity of consumption in the dataset. It is clear from the figure that the consumption patterns of houses are highly diverse, as there are some houses which have very low consumption in summer and some houses with very high consumption in winter.

**[slide]**

The PRECON data includes 26 metadata parameters. Out of several demographic parameters we have chosen to predict the number of people, the number of fans, the number of air conditioners, rooms, refrigerators and the property area. These are some of the important parameters that affect the electricity consumption. The box plot on the right shows the diversity in these parameters. As you can see the number of people living in a single house range from 3 to 11 and the number of air conditioners installed in a house range form 1 to 11. So there is a lot of diversity in the dataset.

**[slide]**

So the core part of our work are the these techniques which model the relationship between monthly electricity consumption and the consumer demographics. One model describes the relationship between 12 months electricity consumption and one of the demographic parameters. These are the modelling techniques that we have applied. We used multivariate linear regression with 12 predictor variable which are the number of months of the whole year, and we use bi-directional selection algorithm to select the predictor variables with significant p-values.

In the support vector regression, we used epsilon regression with radial basis function. In applying the neural network, we have used sigmoid activation function and have tuned the number of layers, neurons and learning rates to achieve the best results.

In all these techniques we have done several iterations to tune the hyperparameters to achieve the minimum error.

**[slide]**

To evaluate the results we have used Mean Absolute Percentage Error in short also known as MAPE. MAPE gives a very good representation of error in the predicted data. It nicely tackles the problem of negation of positive and negative errors because it uses absolute value of error and it also shows the significance of error with respect to the actual value in form of percentage.

**[slide]**

The results that we have achieved so far are very promising. Multivariate linear regression was able to achieve an MAPE 3.5% in the prediction of number of people. This thus shows a strong linear correlation between the electricity consumption in a house and number of people living there. Neural network performed better in prediction of fans, air conditioners and rooms. The property area was not predicted very well and it showed a very high error of 141%. The significance of errors would largely depend on the application.

**[slide]**

The dataset used in this technique includes only 42 households. To achieve better results, the training dataset can be increased in size by including more houses as well as data from different cities. The dataset can also be expanded to include commercial and industrial consumers. The modelling techniques that we have applied can further be explored to better predict the demographics. In linear regression relationship with higher power of predictor variables can be explored. Further advanced techniques such as dimensionality reduction and correlation within the metadata can be explored to achieve better results.