Udacity Capstone Project Proposal
AWS Machine Learning Engineer Nanodegree

Hourly Energy Consumption Forecasting

Domain Background

The focus of my capstone project would be in the time series forecasting domain. This is an area of machine learning/statistical methods that uses past data to forecast future value of a numerical variable. Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. It is a very important problem in many professional fields including finance, sales, electricity, stocks, etc.

Problem Statement

Accurate forecast of future energy consumption in an electrical power grid allows energy providers to balance power generation and consumption in real time to conserve resources at power plants and ensure grid stability. A significant mismatch of power generation and consumption in a power grid could lead to over-frequency or under-frequency in the grid, both of which could result in grid collapse, a term used in electrical engineering to refer to a situation where all generation stations connected to a grid are disconnected from the grid to preserve their health because they cannot supply power to the grid at the current grid frequency. Grid collapse can be very catastrophic for a modern society and often comes with significant financial losses for stake holders.

Dataset and Input

For this project, I will be using a publicly available energy consumption data from 2004 - 2018 available at Kaggle Hourly Energy Consumption provided by PJM Interconnection LLC (PJM). PJM is a regional transmission organization (RTO) in the United States. It is part of the Eastern Interconnection grid operating an electric transmission system serving all or parts of Delaware, Illinois, Indiana, Kentucky, Maryland, Michigan, New Jersey, North Carolina, Ohio, Pennsylvania, Tennessee, Virginia, West Virginia, and the District of Columbia. The hourly power consumption data comes from PJM's website and are in megawatts (MW). The features in the data are described below:

Column Name	Description
Datetime	Datetime in the format yyyy-mm-dd hh:mm:ss
AEP_MW	Megawatt energy consumption

Solution Statement

I intend to explore various machine learning techniques to come up with a model that can accurately predict energy requirement for a day in the future. Steps will include:

1)Data overview and treatment: to make sure data is suitable for modelling
2)Data Analysis and Visualization: this include statistical analysis, time-series decomposition, statistical test.
3)Data pre-processing: train-test split

4)Modelling: First, I'll use the ARIMA techniques to build and test Moving Average (MA) and Autoregressive (AR) models. Using the ARIMA techniques would require caring out Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) analysis. I also intend to tackle this as a linear regression problem using XGBoost or AWS AutoML framework (AutoGluon). This will require performing some feature engineering to extract features from the datetime. I will be examining the feature importance of all predictor variables that go into the regression model.

The final solution would be a well-tuned forecasting model that can provide the most accurate forecast on unseen test data for energy consumption.

Benchmark Model

A simple and commonly used benchmark model for time series forecasting is the simple moving average model. In this method, we forecast the next value(s) in a time series based on the average of a fixed finite number of the previous values. This means that the model does not take into account any trends, seasonality, or other patterns in the data. This is a basic model but it provides a useful benchmark against which other, more advanced models can be compared.

Evaluation Metrics

Before embarking on model building, I need to select which evaluation metrics to assess the benchmark and solution models against. In the context of the domain, the following evaluation metrics will be used to judge the quality of models on unseen test data: *MAE:* Mean absolute error is how far away the forecasts are in comparison to the actual values in the time series. Its measured as the mean of the absolute error between all predictions and corresponding actual value. *RMSE*: Root mean square error also measures how far away the forecasts are in comparison to the actual values, but its the mean of the squares of the error between time-series values and correspond forecast.

Solution Design

Finished project will be completed on **AWS** as described in the steps below:

1)Set Up Amazon Sagemaker Instance with appropriate computing resources
2)Download Dataset from kaggle and upload to amazon S3
3)Extract, transform and load dataset into sagemaker notebook
4)Data analysis and visualization
5)Data pre-processing
6)Modelling and Evaluation
7)Model deployment to sagemaker endpoint