Machine Learning Engineer Nanodegree

Capstone Proposal

Monish Ananthu
September 30th, 2018

Domain Background

SMARD

[Link to SMARD Website (https://www.smard.de/en)](https://www.smard.de/en)

The Bundesnetzagentur's electricity market information platform "SMARD" is an abbreviation of the German term “Strommarktdaten”, which translates to electricity market data. Data that is published on SMARD’s website gives an up-to-date and in-depth overview of what is happening on the German electricity market.

How high is electricity supply and demand? How big is the share of electricity generated from renewable sources? How does electricity consumption change over the course of the day? How much electricity is imported into Germany and exported to its neighbours? SMARD's market data provides answers to these and many other questions.

The following electricity market data categories can be accessed/downloaded:

- Electricity generation
  - Actual generation
  - Forecasted generation
  - Installed capacity
- Electricity consumption
  - Realised consumption
  - Forecasted consumption
- The market
  - Wholesale market price
  - Commercial exchanges
  - Physical flows
- System stability
  - Balancing energy
  - Total costs
  - Primary balancing capacity
  - Secondary balancing capacity
  - Tertiary balancing reserve
  - Exported balancing energy
  - Imported balancing energy

The above data is available from 2015 onwards. The statistical data available is visuallized and limited to a specific subcategory (for example: Electricity generation --> Actual generation). The visualization does not convey how the data is correlated to one another and also the correlation of data between different catagories like "Actual generation" and "Wholesale market price" would be a very interesting to determine.

What makes SMARD Data so interesting?

- Data is already consolidated from different transmission system operators in a standard format
- High frequency of data (in 15 minute / hourly intervals) provides a good basis for data analysis
- Data available from 2015 ist constantly updated

Personal motivations

In my current role, I'm responsible to create new digital services for electromobility. I was curious to know more about the sources of energy produced in Germany and the current trend for regenerative energy. Since in the end, the price is a deciding factor for the end customer I chose to determine the current trend regarding energy prices based on different sources.

Problem Statement

The problem to be solved is the prediction of the wholesale market price of energy [Euro/MWh] using the data available above. The problem at hand is a supervised learning problem in the field of Machine Learning. From the **Datasets and Inputs** section below, we have the following input data:

a. Actual generation
b. Realized Consumption
c. Balancing energy

It is important to find correlations among the above input features and use this information to predict the wholesale market price of energy.

All data is available in CSV format

Optional goals

- Find correlations among available features and discover new insights on the data

Datasets and Inputs

The data sets can be downloaded at [https://www.smard.de/en/downloadcenter/download_market_data (Link)](https://www.smard.de/en/downloadcenter/download_market_data) Select category, sub-category, country = Germany, Dates: 01/01/2015 - 31/12/2015, Filetype: CSV and download file.

We will consider 2 Datesets:

a. Small data set - valid for 2015

b. Large data set - valid from 01/01/2015 - 30/09/2018

The subcategories below refer to feature sets. If a sub-category is not relevant, all features in the feature set can be discarded. Partial relevance means that a part of the the features need to be considered.

Category	Sub-category	Relevant?	Data frequency	File name	Details
Electricity generation	Actual generation	yes	15 mins	DE_Actual generation.csv	Amount of energy generated by different sources at a specific time period
	Forecasted generation	no	15 mins	DE_Prognostizierte Erzeugung.csv	Forecasted features are not relevant
	Installed Capacity	no	NA	DE_Installierte Erzeugungsleistng.csv	Not enough data
Electricity consumption	Realized consumption	yes	15 mins	DE_Actual consumption.csv	Energy consumption at a specific time period
	Forecasted consumption	no	15 mins	DE_Prognostizierter Stromverbrauch.csv	Forecasted features are not relevant
The Market	Wholesale market price	yes (partially)	15 mins	DE_Day-ahead prices.csv	Energy price per MWh. Only data for Germany is relevant
	Commercial exchanges	no	60 mins	DE_Kommerzieller Außenhandel.csv	Energy Imports and Exports are out of scope for price prediction
	Physical flows	no	60 mins	DE_Physikalischer Stromfluss.csv	Energy Imports and Exports are out of scope for price prediction
System stability	Balancing energy	yes	15 mins	DE_Balancing energy.csv	Overall energy balancing volumes and balancing price
	Total costs	no	monthly	DE_Total costs.csv	Not enough data
	Primary balancing capacity	no	15 mins	DE_Primärregelleistung.csv	Energy balancing efforts and resulting costs are not in scope
	Secondary bancing capacity	no	15 mins	DE_Sekundärregelleistung.csv	Energy balancing efforts and resulting costs are not in scope
	Tertiary balancing reserve	no	15 mins	DE_Minutenreserve.csv	Energy balancing efforts and resulting costs are not in scope
	Exported balancing energy	no	NA	NA	No data available for 2015
	Imported balancing energy	no	NA	NA	No data available for 2015

After the initial screening we have the following features:

Category	Sub-category	Feature	Data frequency	Comments
		Date		Date starting 01/01/2015 - 31/12/2015
		Time of day		Timestamps in 15 min intervals for the date range specified above
Electricity generation	Actual generation	Hydropower[MWh]	15 mins	Generated energy in MWh
		Wind offshore[MWh]	15 mins	Generated energy in MWh
		Wind onshore[MWh]	15 mins	Generated energy in MWh
		Photovoltaics[MWh]	15 mins	Generated energy in MWh
		Other renewable[MWh]	15 mins	Generated energy in MWh

Category	Sub-category	Feature	Data frequency	Comments
		Nuclear[MWh]	15 mins	Generated energy in MWh
		Fossil brown coal[MWh]	15 mins	Generated energy in MWh
		Fossil hard coal[MWh]	15 mins	Generated energy in MWh
		Fossil gas[MWh]	15 mins	Generated energy in MWh
		Hydro pumped storage[MWh]	15 mins	Generated energy in MWh
		Other conventional[MWh]	15 mins	Generated energy in MWh
Electricity consumption	Actual consumption	Total[MWh]	15 mins	Feature name needs to be modified to Total consumption for simplicity
The Market	Wholesale market price	Germany/Austria/Luxembourg[Euro/MWh]	60 mins	Market prices of other countries are not relevant and need not be considered. The feature name will be renamed to Price Germany for simplicity
System stability	Balancing energy	Balancing energy volume[MWh]	15 mins	Balancing energy in MWh
		Balancing energy price[Euro/MWh]	15 mins	Price for balancing energy Euro/MWh

Each of the features in the dataset contains a value for a particular time period/interval. The feature we like to predict "Wholesale market price" is available every 60 minutes. This implies that the input features which are currently available every 15 minutes need to be reduced to once every 60 minutes to correspond with the predicted feature.

Why is this data set relevant for the problem?

From the information presented in [2], we see clearly

The electricity market brings supply and demand together.

The main element to control the market is the Price.

We already have supply data i.e. energy supply data from different sources and demand data which is the consumption data. We also have the energy price for any give time period. If supply and demand are key factors which influence the price, we already have the relevant data to analyse using Supervised Machine Learning.

Solution Statement

The prediction of the wholesale market price of energy is a regression problem. We can use regression techniques such as:

1. Linear Regression
2. Polynomial Regression
3. Ridge and Lasso Regression
4. Decision Tree Regression

The mathematical expression for linear regression is

$$y = m_1x_1 + m_2x_2 + m_3x_3 + ... + m_nx_n + b$$

where,
y = prediction
$m_1, m_2, m_3 ... m_n$ are coefficients / slopes
$x_1, x_2, x_3 ... x_n$ are the predictor variables in "n" dimensions
b is the y-intercept

Similarly, the mathematical expression for a polynomial regression is

$$y = m_1 x^3 + m_2 x^2 + m_3 x + m_4$$

For the problem at hand, it is clear that the energy price is dependent on supply and demand. If we're able to establish a relationship (correlate) between supply and demand using one the the regression techniques available, we should be able to determine the how this relationship influences the energy pricing. At the moment it is still unclear, if the regression is linear or polynomial. I hope to find more information when I analyze the data.

Benchmark Model

Electricity is grid-based and very difficult to store. Although batteries can be found in every household, most technologies for the storage of electricity on a large scale are still limited. They are either not fully developed or not profitable on the market. This means that electricity must be produced at the same moment that it is consumed

The electricity market is made up of submarkets with different Prices. There are therefore different trading products on the exchange with different periods of time between purchase and actual supply. Electricity can be traded several years in advance on the futures market and buyers use these long-term contracts to hedge against the risk of rising prices. For this planning certainty, they pay a premium, which sellers then register as additional revenue. Long-term contracts secure income for the producers, which they can use to finance new generating capacity, for example.

As the day of supply draws closer, the actual volumes of consumption and generation can be predicted more accurately, so the short-term spot market consists of two markets with different lead times: the day-ahead and intraday markets. Market players on the day-ahead market trade in electricity for the following day. Bids and offers specifying the amount and supply time must be submitted by midday. The exchange then determines the wholesale price for each hour of the next day and accepts the winning bids and offers. The wholesale price determined in this way is an important reference value for the electricity market, rather like the closing price of a stock on the stock market. For this reason, the day-ahead wholesale prices of the most important electricity exchange EPEX are shown on the SMARD website. Electricity can be traded until 30 minutes before supply in the continuous intraday trading. [2]

There is currently no benchmark model available for the problem at hand. My benchmark model would be using a Linear Regressor in the capstone project to predict the wholesale energy price. This benchmark model will be challenged by other supervised regression techniques to achieve a better result.

Evaluation Metrics

The Evaluation metrics[3] for a Regression based problem are:

1. Mean Absolute Error
2. Mean Squared Error
3. Mean squared logarithmic error
4. Median absolute error

5. R2 Score

I choose R2 Score for the benchmark model and the solution model.

Project Design

The following phases are revelant:

1. **Data Preparation** - Data first needs to be collected and pre-screened for relevance and completeness. The the data can be visualized to check for relationships between different variables and outliers. The data may also need to be normalized and scaled. This data is split into training and testing sets. We will use the majority of the data for training and evaluate model performance later on using the testing set.
2. **Model Selection & Experimentation** - There are many models available for regression. We will try different models in this step and choose the model with the best prediction.
3. **Model Training** - In the data prepraration step we already split the data set into Training and testing set. We will use the training data set to train the model. If necessary we also can use a subset of dat for cross validation, to make sure the model doesn't overfit.
4. **Model Evaluation** - In this step we feed the trained model data it hasn't seen before and evaluate how good the prediction is.
5. **Model Parameter Tuning** - Once we know how a chosen model is performing, we can take a look at the model parameters for possibilities to increase prediction rate. The Grid Search technique is a useful tool to determine the best parameters.
6. **Prediction** - We have chosen a model and tuned it to our problem. This is the final step which helps us fullfill our problem statement.

References

1. https://www.smard.de/en/Ueber_uns/5732 (https://www.smard.de/en/Ueber_uns/5732)
2. https://www.smard.de/en/wiki-article/5884/5840 (https://www.smard.de/en/wiki-article/5884/5840)
3. http://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics (http://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics)
4. https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e (https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e)