report

October 28, 2018

1 Machine Learning Engineer Nanodegree

1.1 Capstone Project

Monish Ananthu October 28th, 2018 ***## Energy Price Prediction***

1.2 I. Definition

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.

The SMARD Website offers real time data for analysis. This data is available for download in different formats. Link to SMARD Website

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

1.2.1 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

1.3 II. Analysis and Preprocessing

The data sets can be downloaded at 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 a Data sets for the years 2015 and 2016.

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	Details
Electricity generation	Actual generation	yes	15 mins	Amount of energy generated by different sources at a specific time period
	Forecasted generation	no	15 mins	Forecasted features are not relevant
	Installed Capacity	no	NA	Not enough data

Category	Sub-category	Relevant?	Data frequency	Details
Electricity consumption	Realized consumption	yes	15 mins	Energy consumption at specific time period
	Forecasted consumption	no	15 mins	Forecasted features are not relevant
The Market	Wholesale market price	yes (partially)	15 mins	Energy price per MWh. Only data for Germany is relevant
	Commercial exchanges	no	60 mins	Energy Imports and Exports are out of scope for price prediction
	Physical flows	no	60 mins	Energy Imports and Exports are out of scope for price prediction

Category	Sub-category	Relevant?	Data frequency	Details
System stability	Balancing energy	yes	15 mins	Overall energy balancing volumes and balancing price
	Total costs	no	monthly	Not enough data
	Primary balancing capacity	no	15 mins	Energy balancing efforts and resulting costs are not in scope
	Secondary bancing capacity	no	15 mins	Energy balancing efforts and resulting costs are not in scope
	Tertiary balancing reserve	no	15 mins	Energy balancing efforts and resulting costs are not in scope

Category	Sub-category	Relevant?	Data frequency	Details
	Exported balancing energy	no	NA	No data available for 2015
	Imported balancing energy	no	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
		Nuclear[MWh]	15 mins	Generated energy in MWh
		Fossil brown coal[MWh]	15 mins	Generated energy in MWh

Category	Sub-category	Feature	Data frequency	Comments
		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.

1.3.1 Data Visualization

We have 4 types of data and we have to explore them separately.

a. Actual generation - contains values of individual energy sources
b. Realized Consumption - total energy consumption (Germany)
c. Balancing energy - Energy required for balancing and price Euro/Mwh
d. Wholesale energy price - target variable

```
In [20]: master_data[actual_generation].describe()
```

```
Out[20]:        biomass_mwh  hydropower_mwh  wind_offshore_mwh  wind_onshore_mwh  \
         count  17393.000000    17461.000000       17483.000000      17479.000000
         mean    1059.656126      453.039030         285.107447       1907.624707
         std       89.666552      128.472593         231.897884       1606.903028
         min      660.500000      202.250000           0.000000         28.000000
         25%      986.250000      354.250000          85.750000        726.250000
         50%     1080.250000      423.750000         210.250000       1412.250000
         75%     1147.000000      530.750000         481.750000       2616.250000
         max     1206.000000      785.250000         915.500000       7738.250000

               photovoltaics_mwh  other_renewable_mwh   nuclear_mwh  \
         count       17450.000000         17384.000000  17500.000000
```

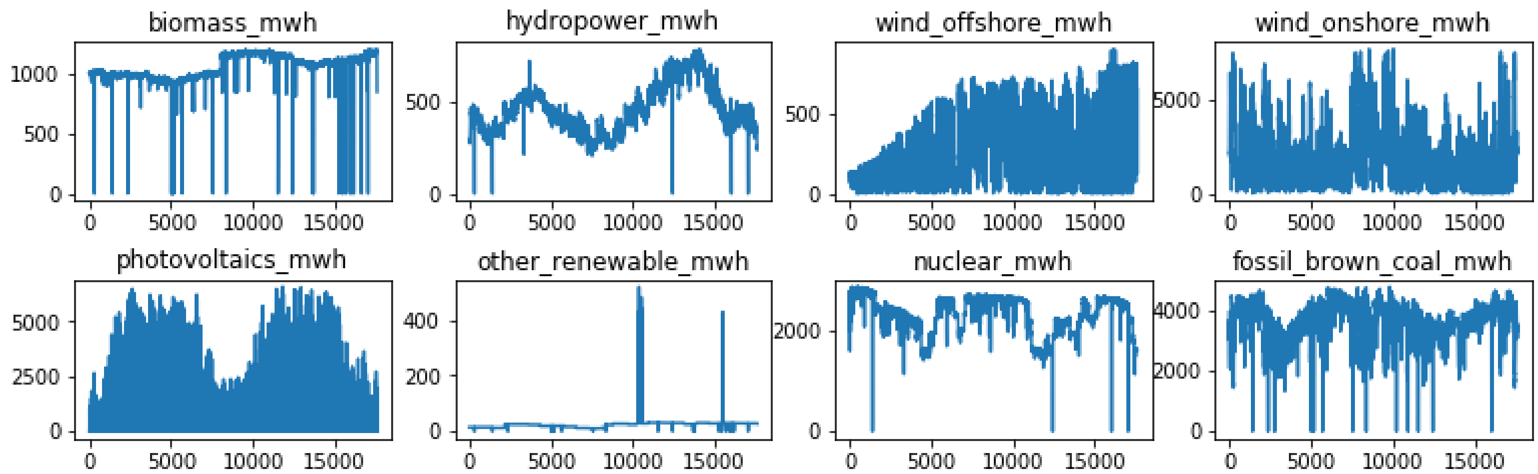
```
mean          985.667722            23.507248   2345.017114
std          1511.748379            15.589565    360.398637
min             0.000000             6.500000   1130.500000
25%             0.000000            16.250000   2156.250000
50%            17.250000            25.000000   2457.000000
75%          1579.000000            27.750000   2630.250000
max          6563.000000           518.250000   2868.500000

       fossil_brown_coal_mwh  fossil_hard_coal_mwh  fossil_gas_mwh  \
count           17400.000000          17428.000000    17446.000000
mean             3762.063707           2337.405899      362.094334
std               576.352412           1231.525012      311.543847
min              1332.500000            103.750000        7.250000
25%              3453.000000           1250.062500      153.312500
50%              3836.750000           2393.125000      235.125000
75%              4186.500000           3350.562500      455.187500
max              4807.500000           5169.500000     2364.750000

       hydro_pumped_storage_mwh  other_conventional_mwh
count              17475.000000            16579.000000
mean                 175.073462             1437.941764
std                  226.936886              609.139779
min                    0.000000              174.750000
25%                    1.250000             1021.750000
50%                   74.500000             1334.500000
75%                  275.000000             1716.750000
max                 1598.250000             8127.500000
```

1. Detecting yearly patterns As the data is spread over 2 years, we visualize the data to discover yearly patterns.

In [22]:
```
# Plot actual generation
fig = plt.figure(figsize=(13,8))
fig.subplots_adjust(hspace=.5)

plt.subplot(4,4,1)
```

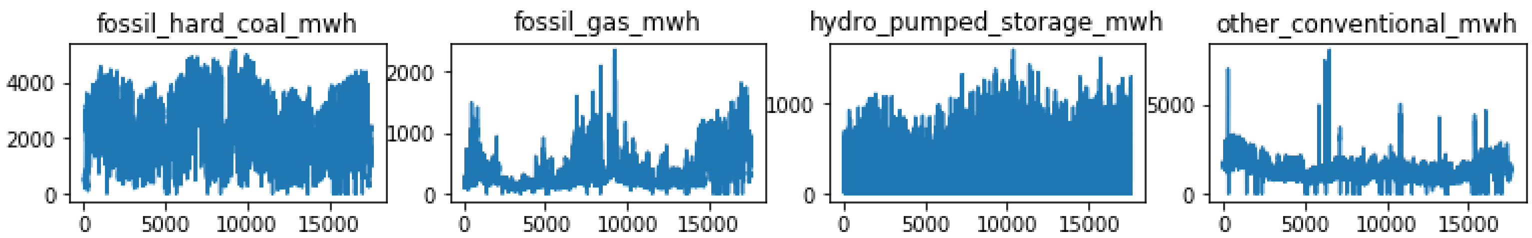
```
master_data['biomass_mwh'].plot()
plt.title('biomass_mwh')

plt.subplot(4,4,2)
master_data['hydropower_mwh'].plot()
plt.title('hydropower_mwh')

plt.subplot(4,4,3)
master_data['wind_offshore_mwh'].plot()
plt.title('wind_offshore_mwh')

plt.subplot(4,4,4)
master_data['wind_onshore_mwh'].plot()
plt.title('wind_onshore_mwh')

plt.subplot(4,4,5)
master_data['photovoltaics_mwh'].plot()
plt.title('photovoltaics_mwh')

plt.subplot(4,4,6)
master_data['other_renewable_mwh'].plot()
plt.title('other_renewable_mwh')

plt.subplot(4,4,7)
master_data['nuclear_mwh'].plot()
plt.title('nuclear_mwh')

plt.subplot(4,4,8)
master_data['fossil_brown_coal_mwh'].plot()
plt.title('fossil_brown_coal_mwh')
```

```
Out[22]: Text(0.5,1,'fossil_brown_coal_mwh')
```




```
In [23]: fig = plt.figure(figsize=(13,8))
         fig.subplots_adjust(hspace=.5)

         plt.subplot(4,4,9)
         master_data['fossil_hard_coal_mwh'].plot()
         plt.title('fossil_hard_coal_mwh')

         plt.subplot(4,4,10)
         master_data['fossil_gas_mwh'].plot()
         plt.title('fossil_gas_mwh')

         plt.subplot(4,4,11)
         master_data['hydro_pumped_storage_mwh'].plot()
         plt.title('hydro_pumped_storage_mwh')

         plt.subplot(4,4,12)
         master_data['other_conventional_mwh'].plot()
         plt.title('other_conventional_mwh')
```

Out[23]: Text(0.5,1,'other_conventional_mwh')



We currently have 17596 hourly dataset rows over a persiod of 2 years 2015-2016 and 2016-2017 i.e. 8798 per year.

Looking at the above data we can observe that non-renewable sources of energy have a pattern which can be predicted if we were to divide the graphs right in the middle. This is true for nuclear, fossil brown coal, fossil hard coal, fossil gas, hydro pumped storage and other conventional energy sources.

The same is not consistent for renewable sources. Wind energy is weather dependent which does not have set patterns. If we observe biomass and hydropower, they reflect similar patterns but in varying sizes. This could be a result of increasing yearly investments in these energy sources. Photovoltaics appears to be consistent with a predictable pattern.

In [24]:
```
master_data[actual_consumption].describe()
```

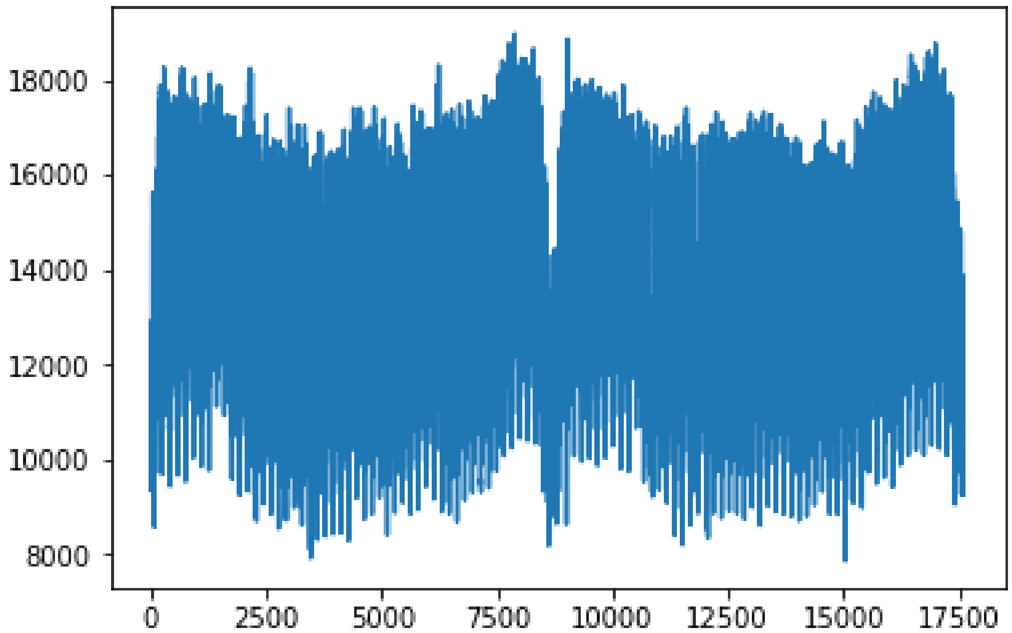
Out[24]:

	total_consumption_mwh
count	17596.000000
mean	13663.896255
std	2477.577389
min	7856.750000
25%	11586.750000
50%	13582.500000
75%	15906.437500
max	18975.250000

The mean and median of 'total_consumption_mwh' varies to a large extent when compared to features grouped under 'actual_generation'. Scaling is therefore necessary.

In [25]:
```
master_data['total_consumption_mwh'].plot()
```

```
Out[25]: <matplotlib.axes._subplots.AxesSubplot at 0xa8efb70>
```



The consumption data for the second year seems to follow a similar pattern as the first year. This implies that the energy consumption for a certain time of year is consistent to the previous year.

```
In [26]: master_data[balancing_energy].describe()
```

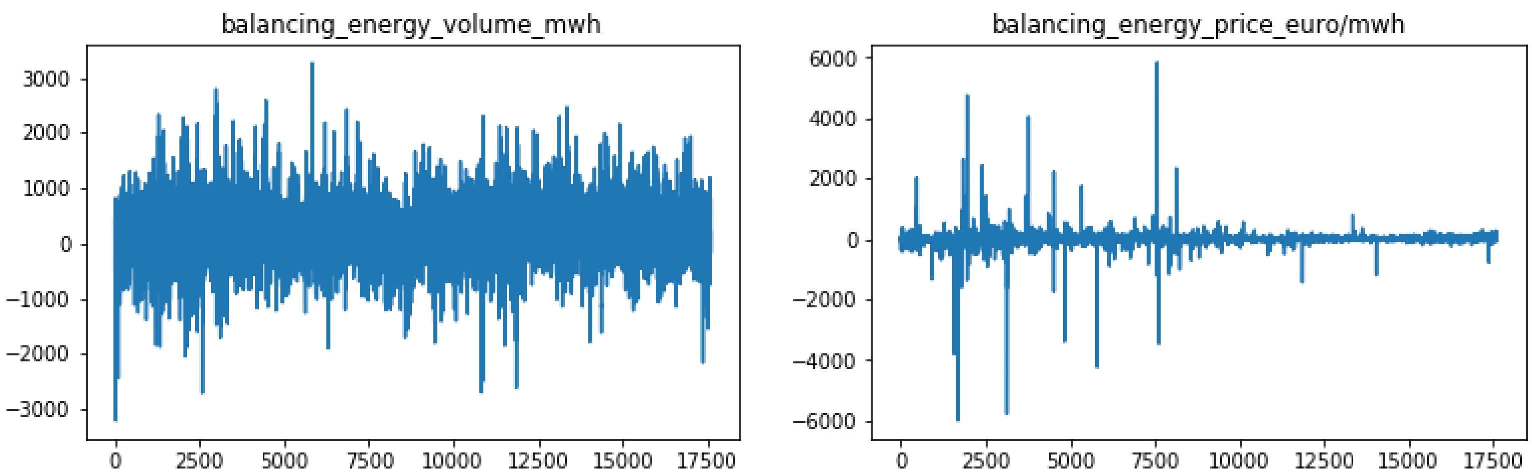
```
Out[26]:        balancing_energy_volume_mwh  balancing_energy_price_euro/mwh
         count                 17596.000000                     17596.000000
         mean                    165.372471                        31.917856
         std                     464.695247                       150.874693
         min                   -3210.000000                     -5997.420000
         25%                    -102.250000                         1.890000
```

```
50%                160.000000              39.940000
75%                437.000000              60.882500
max               3271.000000            5824.630000
```

Similar to 'total_consumption_mwh', Data Scaling is also required.

In [27]:
```python
fig = plt.figure(figsize=(13,8))

plt.subplot(2,2,1)
master_data['balancing_energy_volume_mwh'].plot()
plt.title('balancing_energy_volume_mwh')

plt.subplot(2,2,2)
master_data['balancing_energy_price_euro/mwh'].plot()
plt.title('balancing_energy_price_euro/mwh')
```

Out[27]:
```
Text(0.5,1,'balancing_energy_price_euro/mwh')
```



The balancing energy volume contains a consistent yearly pattern. This does not apply to the balancing energy price which appears more random.

```
In [28]: master_data[target].describe()
```

```
Out[28]:        price_germany_euro/mwh
         count             17476.000000
         mean                 30.386641
         std                  12.551637
         min                -130.090000
         25%                  23.460000
         50%                  29.660000
         75%                  37.060000
         max                 104.960000
```

We still have 120 missing values for **price_germany**. The reason for missing values could be an error in the logging system. In other cases if we look at the data above, we notice negative values. In certain rare periods particularly at the end of the year, there is a such a surplus of wind energy produced that the prices go below zero i.e the consumer gets paid for consuming energy. In the energy market, there is always supply and demand and hence, a price for each time period. Earlier for energy generation we replaced all missing values with zeroes. The case here is different and a price of zero has a false implication and would have a negative influence on the prediction.

The best strategy here would be to either use the Median or the mean.

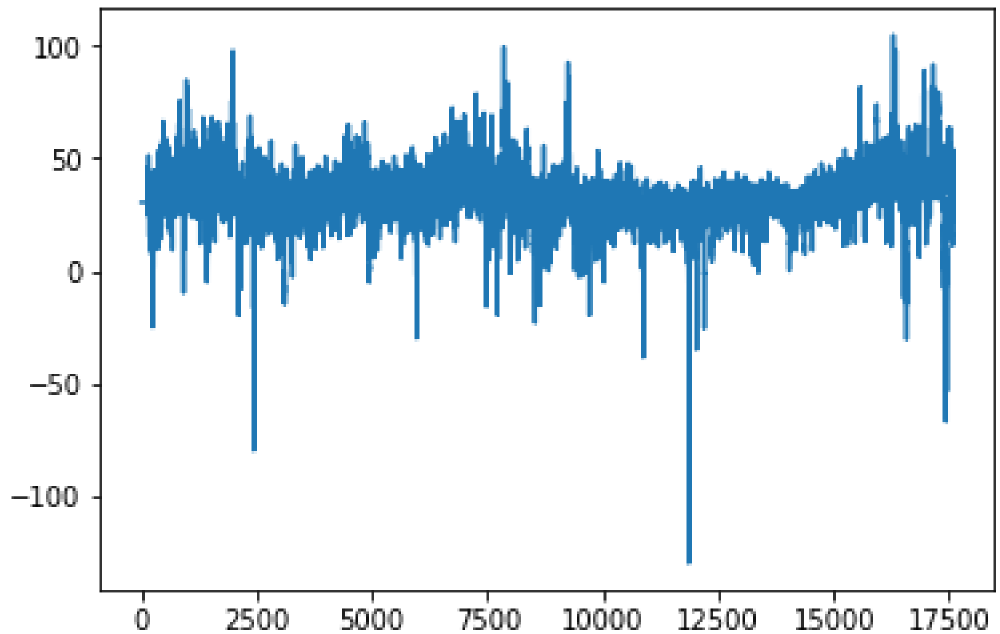
Median = 29.66 Mean = 30.386641

Since there isn't a large difference between Median and Mean, let us consider the Mean value to fill the missing values

```
In [29]: master_data['price_germany_euro/mwh'].fillna(master_data['price_germany_euro/mwh']
                                                           .mean(),inplace=True)
```

```
In [30]: master_data['price_germany_euro/mwh'].plot()
```

```
Out[30]: <matplotlib.axes._subplots.AxesSubplot at 0xa94cf28>
```



From the above figure we can observe a pattern with a lot of activity at start and end of the year. As the year proceeds, it appears more constant. The pattern cannot be reproduced the next year 100% but the tendency and value ranges remain the same. Moreover, pricing is a complex issue which not only depends on demand and supply but other factors which we haven't considered in our case. One factor could be the increase of investment every year in renewable energies increasing output but also at the same time making renewable energy expensive in the initial phases.

Since the machine learning models cannot handle timeseries data, we will delete time related data like 'date' and 'time_of_day', however perserving the order of recorded data as the order of data is essential for training a supervised regression model for time series.

Our final list of features can be seen below.

```
In [32]: list(master_data)

Out[32]: ['biomass_mwh',
          'hydropower_mwh',
```

```
'wind_offshore_mwh',
'wind_onshore_mwh',
'photovoltaics_mwh',
'other_renewable_mwh',
'nuclear_mwh',
'fossil_brown_coal_mwh',
'fossil_hard_coal_mwh',
'fossil_gas_mwh',
'hydro_pumped_storage_mwh',
'other_conventional_mwh',
'total_consumption_mwh',
'balancing_energy_volume_mwh',
'balancing_energy_price_euro/mwh',
'price_germany_euro/mwh']
```

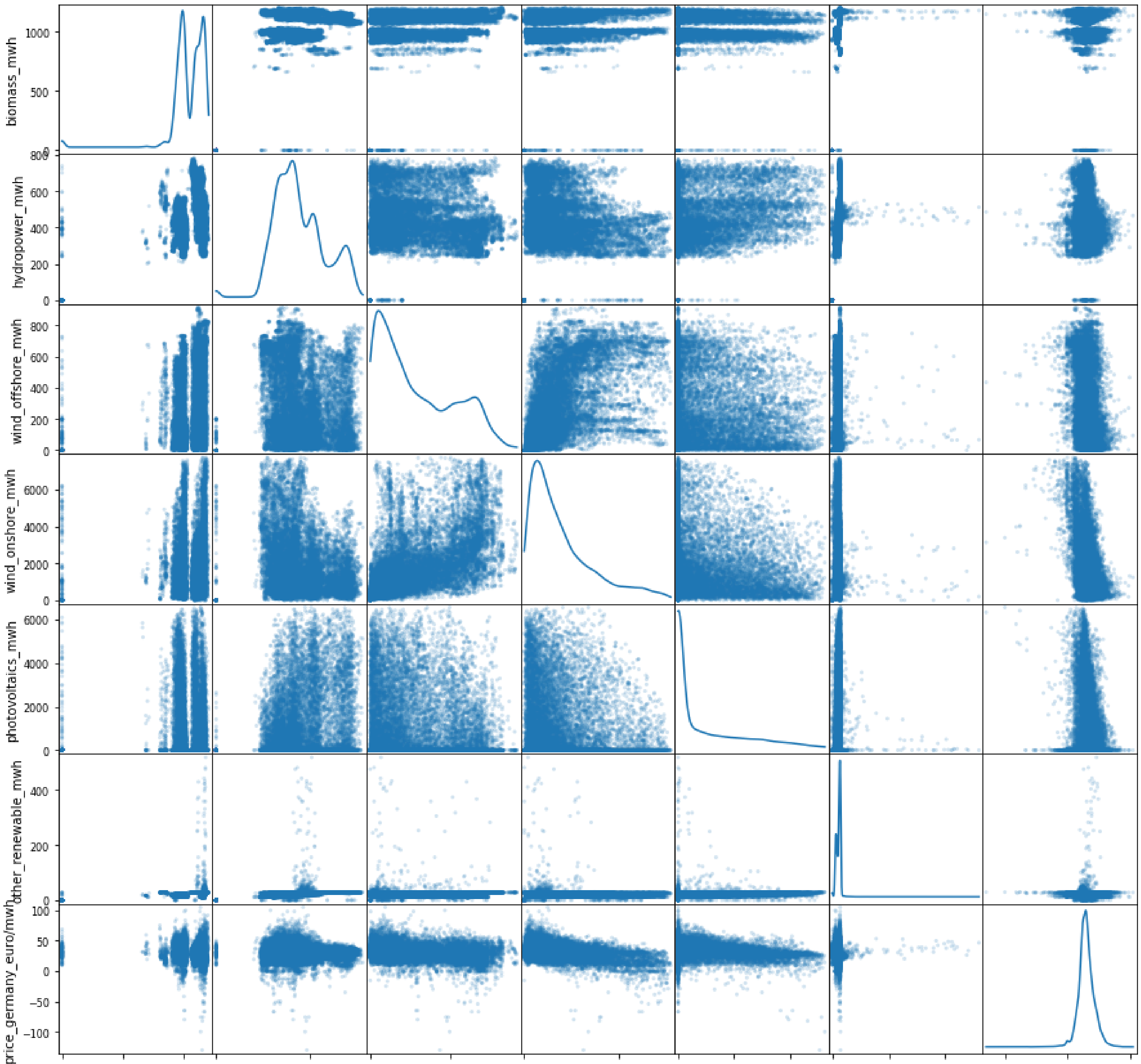
2. Linear correlation among input features and target variable In the scatter plots below, we attempt to find linear correlations between the input features and the target variable.


```
In [33]: from pandas.plotting import scatter_matrix
         scatter_matrix(master_data[['biomass_mwh','hydropower_mwh','wind_offshore_mwh',
                                     'wind_onshore_mwh','photovoltaics_mwh','other_renewable_mwh'
                                     ,'price_germany_euro/mwh']], alpha=0.2, figsize=(15,15), diagonal='kde')
```

```
Out[33]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x000000000ACE81D0>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000AD18940>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000AD3EF28>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000AD6B5F8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000AD92C88>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000AD92CC0>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000ADEC9E8>],
                [<matplotlib.axes._subplots.AxesSubplot object at 0x000000000AE1F0B8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000AE46748>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000AE6FDD8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000B1804A8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000B1A9B38>,
```

```
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000B1DB208>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000B204898>],
[<matplotlib.axes._subplots.AxesSubplot object at 0x000000000B22AF28>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000B25D5F8>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000B285C88>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000B2B5358>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000B2DE9E8>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000B3100B8>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000B337748>],
[<matplotlib.axes._subplots.AxesSubplot object at 0x000000000B3D1DD8>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000B4044A8>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000B42AB38>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000C56F208>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000C703898>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000C72EF28>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000CBFF5F8>],
[<matplotlib.axes._subplots.AxesSubplot object at 0x000000000CC26C88>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000CCB9358>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000CCE09E8>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000CD120B8>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000D229748>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000D252DD8>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000D2854A8>],
[<matplotlib.axes._subplots.AxesSubplot object at 0x000000000D2ADB38>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000D2DF208>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000D308898>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000D37FF28>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000D3AF5F8>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000D41AC88>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000D44B358>],
[<matplotlib.axes._subplots.AxesSubplot object at 0x000000000D4729E8>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000D4A40B8>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000D4CE748>,
 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000D4F3DD8>,
```

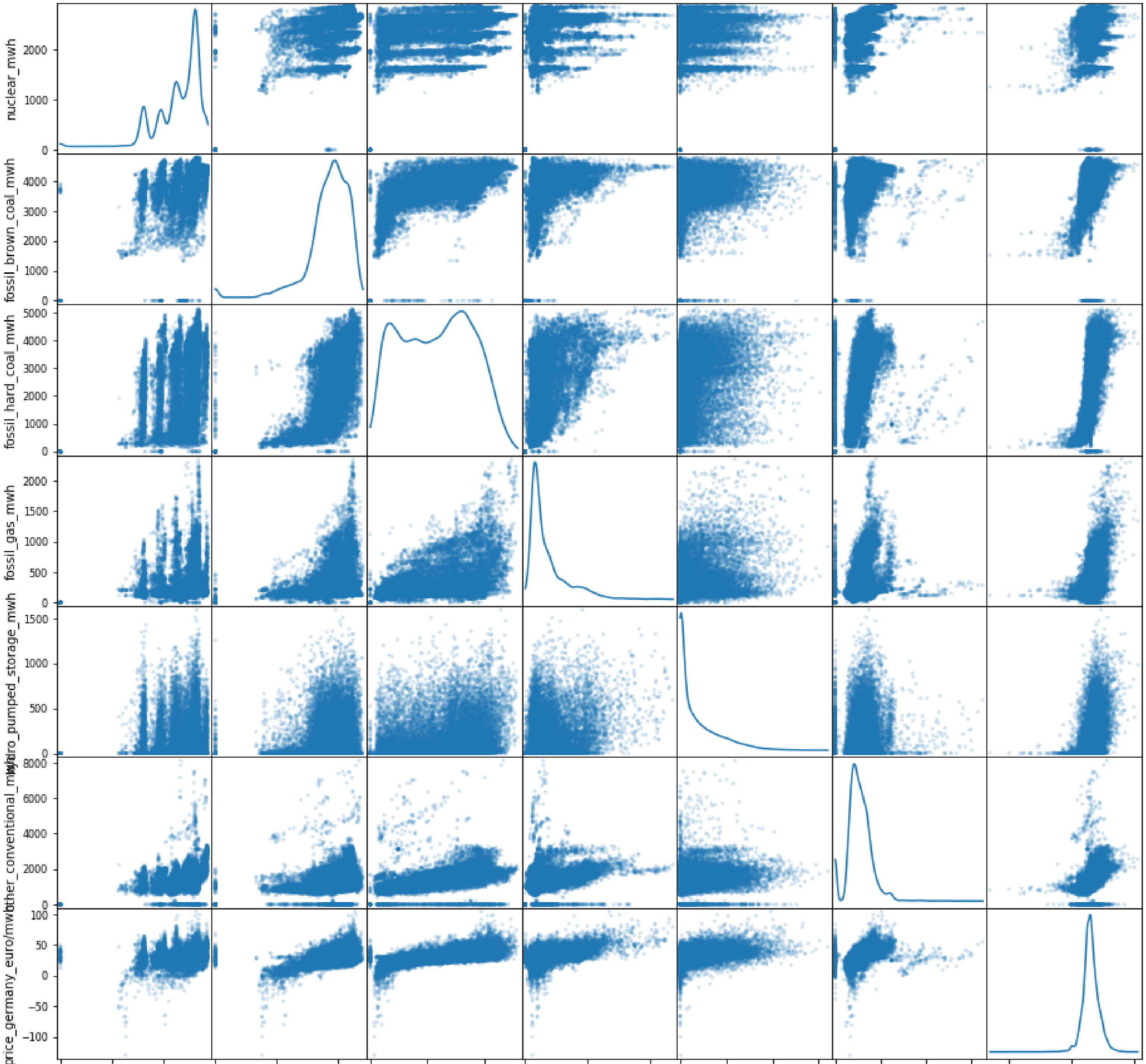
```
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000D5264A8>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000D54FB38>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000D582208>]],
dtype=object)
```


biomass_mwh
0
500
1000
hydropower_mwh
0
200
400
600
800
wind_offshore_mwh
0
200
400
600
800
wind_onshore_mwh
0
2000
4000
6000
photovoltaics_mwh
0
2000
4000
6000
other_renewable_mwh
0
200
400
price_germany_euro/mwh
−100
−50
0
50
100

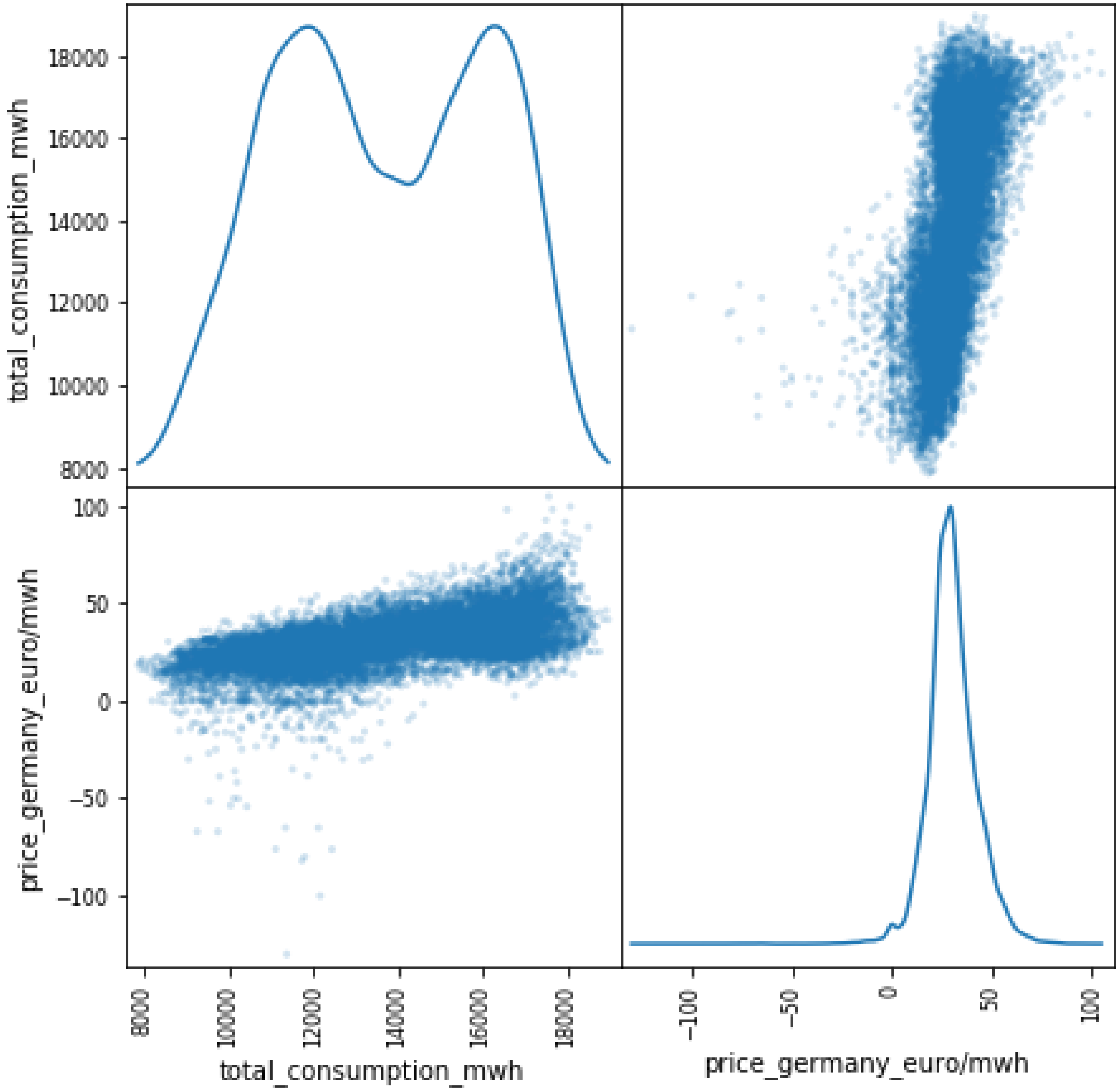
```
In [34]: scatter_matrix(master_data[['nuclear_mwh','fossil_brown_coal_mwh','fossil_hard_coal_mwh',
                                  'fossil_gas_mwh','hydro_pumped_storage_mwh', 'other_conventional_mwh',
                                         'price_germany_euro/mwh']], alpha=0.2, figsize=(15.5,15.5), diagonal='kde')
```

```
Out[34]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x000000000CAD04E0>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000CABB5C0>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000CAF4A90>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000CB180F0>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000CB3F780>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000CB3F7B8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000CB984E0>],
                [<matplotlib.axes._subplots.AxesSubplot object at 0x000000000DA6EB38>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000DAA1208>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000DAC9898>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000DAF3F28>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000DB235F8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000DB4CC88>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000DFAE358>],
                [<matplotlib.axes._subplots.AxesSubplot object at 0x000000000DFD59E8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E0070B8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E02F748>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E058DD8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E08A4A8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E0B0B38>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E526208>],
                [<matplotlib.axes._subplots.AxesSubplot object at 0x000000000E54A898>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E574F28>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E5A75F8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E5CFC88>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E600358>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E62C9E8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E65C0B8>],
```

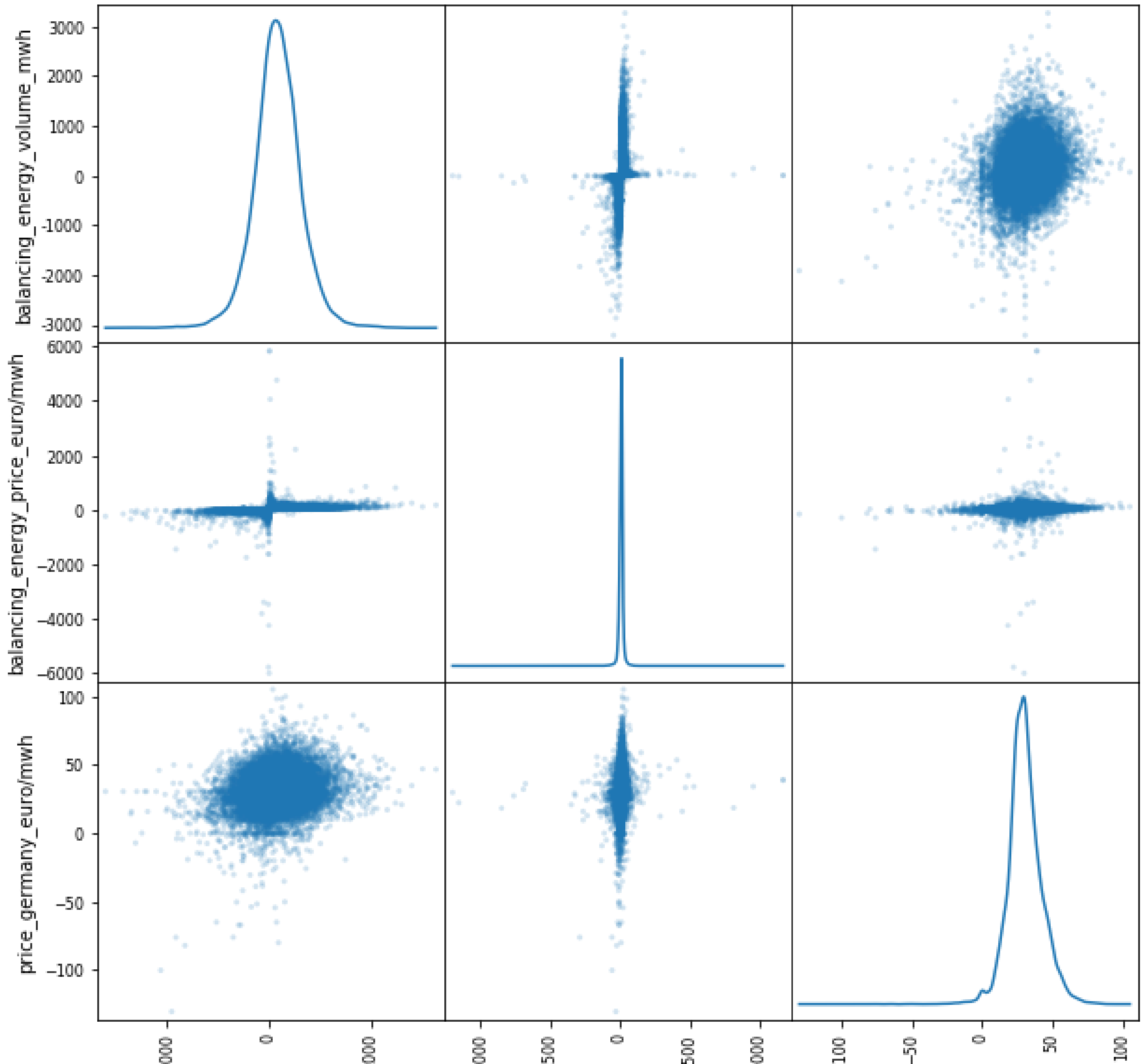
```
 [<matplotlib.axes._subplots.AxesSubplot object at 0x000000000E683748>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E6AADD8>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E8FC4A8>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E925B38>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000E955208>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000ECAE898>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000ECD5F28>],
 [<matplotlib.axes._subplots.AxesSubplot object at 0x000000000ED085F8>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000ED32C88>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000ED63358>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000ED8A9E8>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000EDBB0B8>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000EDE3748>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000EE0BDD8>],
 [<matplotlib.axes._subplots.AxesSubplot object at 0x000000000EE404A8>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000FE67B38>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000FE98208>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000FEC1898>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000FEE9F28>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000FF1A5F8>,
  <matplotlib.axes._subplots.AxesSubplot object at 0x000000000FF44C88>]],
dtype=object)
```


nuclear_mwh
fossil_brown_coal_mwh
fossil_hard_coal_mwh
fossil_gas_mwh
hydro_pumped_storage_mwh
other_conventional_mwh
price_germany_euro/mwh

```
In [35]: scatter_matrix(master_data[['total_consumption_mwh','price_germany_euro/mwh']],
                        alpha=0.2, figsize=(7,7), diagonal='kde')

Out[35]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x00000000106F5860>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x00000000107BF860>],
                [<matplotlib.axes._subplots.AxesSubplot object at 0x0000000010B2F7B8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x0000000010B50AC8>]],
               dtype=object)
```


total_consumption_mwh
18000
16000
14000
12000
10000
8000
price_germany_euro/mwh
100
50
0
−50
−100
8000
100000
120000
140000
160000
180000
total_consumption_mwh
−100
−50
0
50
100
price_germany_euro/mwh

```
In [36]: scatter_matrix(master_data[['balancing_energy_volume_mwh','balancing_energy_price_euro/mwh',
                                     'price_germany_euro/mwh']], alpha=0.2, figsize=(10,10), diagonal='kde')

Out[36]: array([[<matplotlib.axes._subplots.AxesSubplot object at 0x0000000010BCBB70>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x0000000010CB67B8>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x0000000010CDEE10>],
                [<matplotlib.axes._subplots.AxesSubplot object at 0x0000000010D104E0>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x0000000010D38B70>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x0000000010D38BA8>],
                [<matplotlib.axes._subplots.AxesSubplot object at 0x0000000010D918D0>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x0000000010DBBF60>,
                 <matplotlib.axes._subplots.AxesSubplot object at 0x0000000010DEA630>]],
               dtype=object)
```


balancing_energy_volume_mwh
balancing_energy_price_euro/mwh
price_germany_euro/mwh

As we can observe we cannot identify a direct linear correlation between the input features and the target variable.

1.3.2 Data Preparation

For data preparation / Preprocessing we follow the following steps:

Data Structuring 1. Features which have data entries every 15 minutes (Actual generation, actual consumption, balancing energy) need to be reduced to a data entry every 60 minutes (hourly) 2. Features from 1 will be merged with other features (wholesale market price) which already are on an hourly basis 3. Drop non relevant features 4. Simplify feature names for better readability 5. Drop date and time_of_day time series data as it cannot be handled by machine learning algorithms

Dealing with NaNs Energy generation data with NaNs imply that there was no energy produced for a specified time period. These entries can be filled with zeros Target variable price_germany with NaNs should be treated differently compared to energy generation data. Here we will replace NaNs with mean / median as we will see later

Outlier Detection and Elimination Using Tukey's method of Outlier detection [1], we look for datasets which are 1.5 times below the 1st quartile (25th percentile) and 1.5 times above the 3rd quartile (75th percentile) and delete them from the master dataset. Tukey's method detects 5326 outliers, which will be removed.

Feature Scaling From our data analysis we observed that each of the features are on a different scale. There are significant differences in the mean values of each of the individual features. We cannot use this data to train the machine learning model without scaling them. In our case, we will use the MinMaxScaler to scale all the features as some features also include negative values. The values will be scaled between -1 and 1 [5].

Avoiding Look ahead bias and cross validation for the energy Time Series data set The available data set is a time series data set on an hourly basis. Regular cross validation methods like kfold, holdout are not appropriate due to the look ahead bias they generate. Considering the current time 't', we are predicting data for a future time period 't + n', where x is the difference in time to reach the target time for our prediction. Here it is important to not use future data for out training set. We need a special function 'TimeSeriesSplit' to generate training and testing data sets without look forward bias. 'TimeSeriesSplit' ensures that the future data is always used as the test set for prediction to simulate learning under real conditions.

1.3.3 Benchmark Model

Since energy price prediction is a classic regression problem, we will start with Linear Regression as our benchmark model.

```
In [41]: # import relevant libraries
         from sklearn.model_selection import TimeSeriesSplit
         from sklearn import metrics, preprocessing
```

```
from sklearn.preprocessing import MinMaxScaler
import time
```

```
In [42]: from sklearn.linear_model import LinearRegression, Ridge, Lasso
         # Perform Linear Regression using TimeSeriesSplit
         l_tscv = TimeSeriesSplit(n_splits=200)
         l_cv = l_tscv.split(X)

         l_reg = LinearRegression()
         train_list = []
         test_list = []
         rmse_list = []

         # set start time
         start = time.time()

         # generate train and test indices and train each set
         for train_index, test_index in l_cv:
             #print("TRAIN:", train_index, "TEST:", test_index)
             X_train, X_test = X.iloc[train_index], X.iloc[test_index]
             y_train, y_test = y.iloc[train_index], y.iloc[test_index]
             l_reg.fit(X_train,y_train)
             y_pred = l_reg.predict(X_test)
             train_score = l_reg.score(X_train,y_train)
             test_score = l_reg.score(X_test,y_test)
             train_list.append(train_score)
             test_list.append(test_score)
             rmse = np.sqrt(metrics.mean_squared_error(y_test, y_pred))
             rmse_list.append(rmse)

         # set end time
         end = time.time()

         print("Average RMSE Score = ",np.mean(rmse_list))
         print("Average Train Prediction Accuracy =",np.mean(train_list))
```

```
    print("Average Test Prediction Accuracy =",np.mean(test_list))
    print("Execution time: ",end - start)
```

```
Average RMSE Score =  4.953005879681328
Average Train Prediction Accuracy = 0.758279100701416
Average Test Prediction Accuracy = 0.4243483359932581
Execution time:  0.7720000743865967
```

The Linear Regression benchmark model has a test prediction accuray of **42.43%**

1.4 III. Methodology

1.4.1 Considered Regression Models for prediction

Regularized versions of Linear Regression

1) Ridge
2) Lasso

Ensemble models

3) Random Forest Regression
4) Gradient Boosting Regression

Tree Based Gradient Boosting

5) Light GBM Regression

Gradient Boosted Decision Tree

6) XGBoost Regression

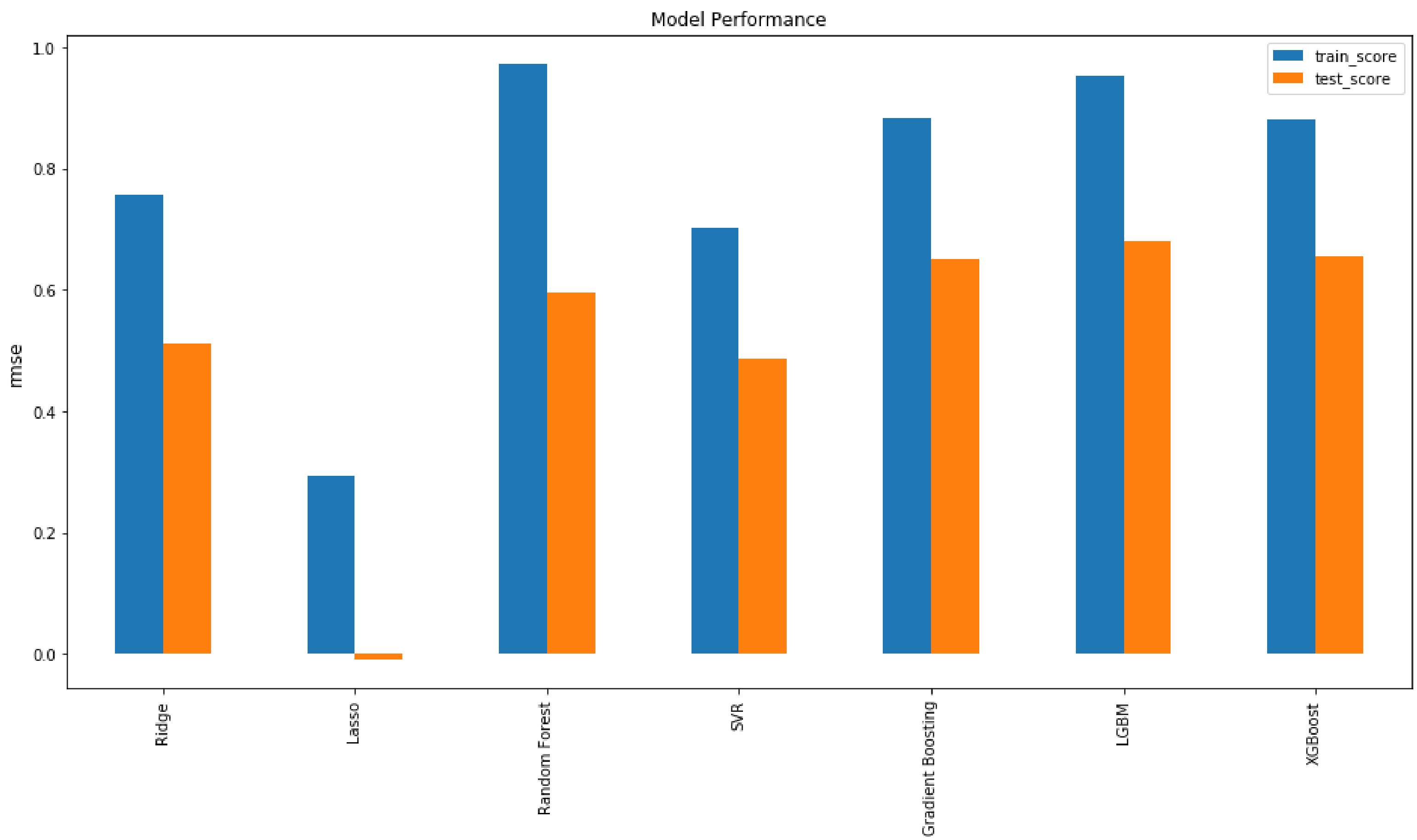
After running different regression models on the data set, the results are summarized below. Let us compare each model based on the following criteria: - Model Performance Train Score vs Test Score - Training time - Root Mean Square Error (RMSE) value

In [46]:
```
# Construct dataframe from dictionary list
model_summary = pd.DataFrame.from_dict(perf_list)
model_summary.rename(index={0:'Ridge',1:'Lasso', 2:'Random Forest',
                            3:'SVR', 4:'Gradient Boosting', 5:'LGBM',
                            6:'XGBoost'}, inplace=True)
model_summary
```

Out[46]:
```
                                        name      rmse  test_score  \
Ridge                                  Ridge  4.976745    0.511779
Lasso                                  Lasso  5.529965   -0.008370
Random Forest          RandomForestRegressor  5.401900    0.596863
SVR                                      SVR  5.387532    0.486393
Gradient Boosting  GradientBoostingRegressor  5.260943    0.649664
LGBM                           LGBMRegressor  5.150525    0.679183
XGBoost                         XGBRegressor  5.070604    0.655381

                   train_score  train_time
Ridge                 0.758339       0.268
Lasso                 0.292662       0.260
Random Forest         0.972212      39.046
SVR                   0.701419     259.357
Gradient Boosting     0.883455      40.815
LGBM                  0.953913      20.458
XGBoost               0.882269      33.593
```

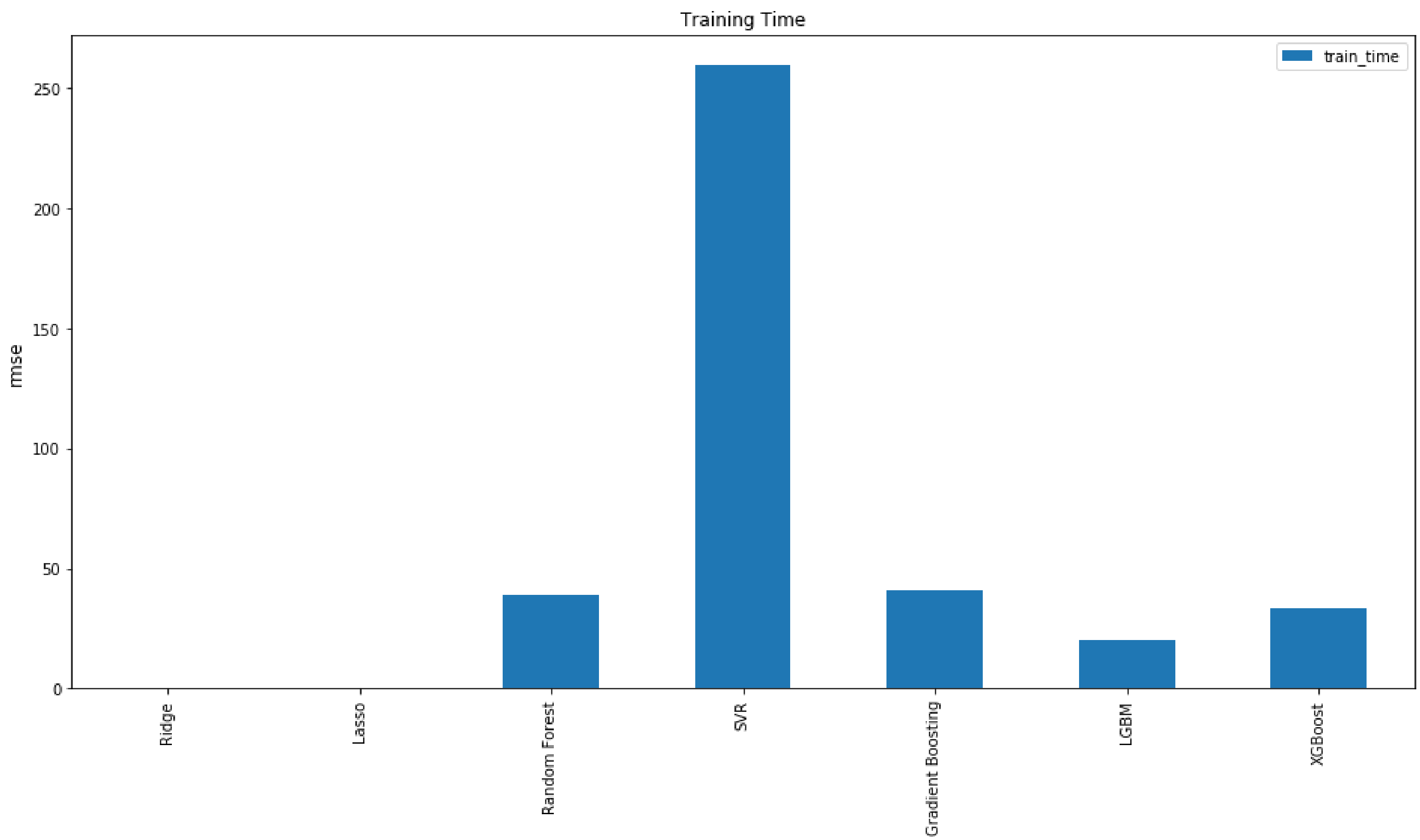
In [47]:
```
# Plot train and test scores
axis = model_summary[["train_score", "test_score"]].plot(kind="bar",
                      title="Model Performance", figsize=(16, 8))
axis.set_ylabel("rmse", fontsize="large")
```

Out[47]:
```
Text(0,0.5,'rmse')
```



The best model according to test scores is LGBM. Although XGBoost and Gradient Boosting Regressor perform quite well.

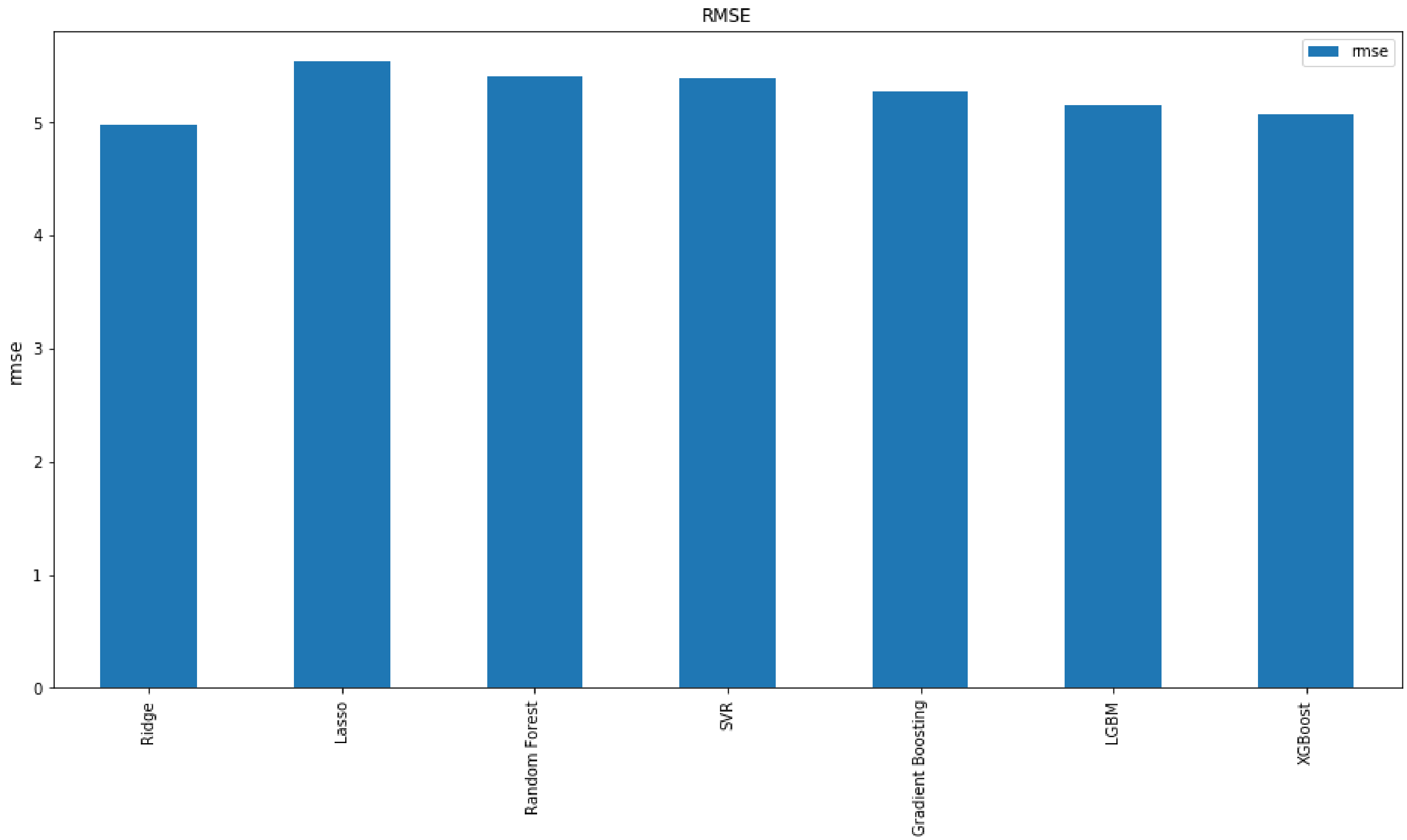
```
In [48]: # Plot train and test scores
         axis = model_summary[["train_time"]].plot(kind="bar", title="Training Time", figsize=(16, 8))
         axis.set_ylabel("rmse", fontsize="large")

Out[48]: Text(0,0.5,'rmse')
```



LGBM doesn't only have the best accuracy, but also has the least training time. Accuracy and training time are 2 important factors if we consider to deploy this model in a productive environment.

```python
In [49]: # Plot train and test scores
         axis = model_summary[["rmse"]].plot(kind="bar", title="RMSE", figsize=(16, 8))
         axis.set_ylabel("rmse", fontsize="large")
```

```
Out[49]: Text(0,0.5,'rmse')
```



The RMSE Scores for all considered models are on a similar level and LGBM has a lower RMSE compared to most of the models.
Best Model : LGBM Regression **Worst Model**: Lasso

1.4.2 Hyper Parameter Tuning

The hyper parameters of LGBM Regressor can now be tuned using GridSearch CV for better fitting leading to higher accuracy of prediction.

Accuracy: 67.91% Error rate (RMSE): 4.36

```
In [50]: lgbm = LGBMRegressor(min_data=1)
         lgbm_param = {'min_data': [1],'min_data_in_bin':[1],'num_leaves': [50,60],
                       'min_data_in_leaf': [50,60,70],'max_depth': [30,50,70]}
         tscv = TimeSeriesSplit(n_splits=70)
         cv = tscv.split(X)
         lgbm_cv = GridSearchCV(lgbm, lgbm_param, cv=cv, verbose=1)
         lgbm_cv.fit(X,y)
         lgbm_cv.best_params_
         lgbm_result = lgbm_cv.cv_results_
         print("Best LGBM Parameters :",lgbm_cv.best_params_)
         print("LGB Score = ",lgbm_cv.best_score_)
```



```
Fitting 70 folds for each of 18 candidates, totalling 1260 fits


[Parallel(n_jobs=1)]: Done 1260 out of 1260 | elapsed:  7.2min finished


Best LGBM Parameters : {'max_depth': 30, 'min_data': 1, 'min_data_in_bin': 1, 'min_data_in_leaf': 60, 'num_leaves': 50}
LGB Score =  0.682210714016851
```

1.5 Reflection

Based on the results obtained from real world energy data, an accuracy of 68.22% is a good result when we take into account the factors that we haven't accounted for in our problem.

Hypothesis 1 We have only considered the energy produced and consumed inside Germany. Germany also exports and imports energy from its neighbours. These factors have not been accounted for in this dataset. If the exports or import needs have changed between 2015 and 2016, they have not been considered.

Hypothesis 2 Some of the energy generation features are dependent on weather, for example wind or water. We have to calculate certain amount of loss with these features.

Hypothesis 3 The information available does not account for the local energy distribution within Germany and the network structures. If a certain amount of energy is available at a certain point of time, this does not guarantee that the energy can be supplied throughout Germany when demand arises.

For any given market, supply and demand define the price. The energy data that was analyzed shows the complexity of an energy market and multiple dependencies that determine prices. Therefore a greater increase in prediction accuracy can only be achieved when we increase the complexity of the current model.

1.5.1 References

[1] http://colingorrie.github.io/outlier-detection.html (Tukey's Method for outlier detection) [2] https://medium.com/apteo/avoid-time-loops-with-cross-validation-aa595318543e (Look ahead bias) [3] https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/ (XGBoost) [4] https://smard.de [5] http://benalexkeen.com/feature-scaling-with-scikit-learn/ (MinMaxScaler)