Delaware Hackathon: Basetable Report

Group 8: Patrick Dundon, Ajay Parihar, Manjunagaraj Rudrappa, Shivam Sarin

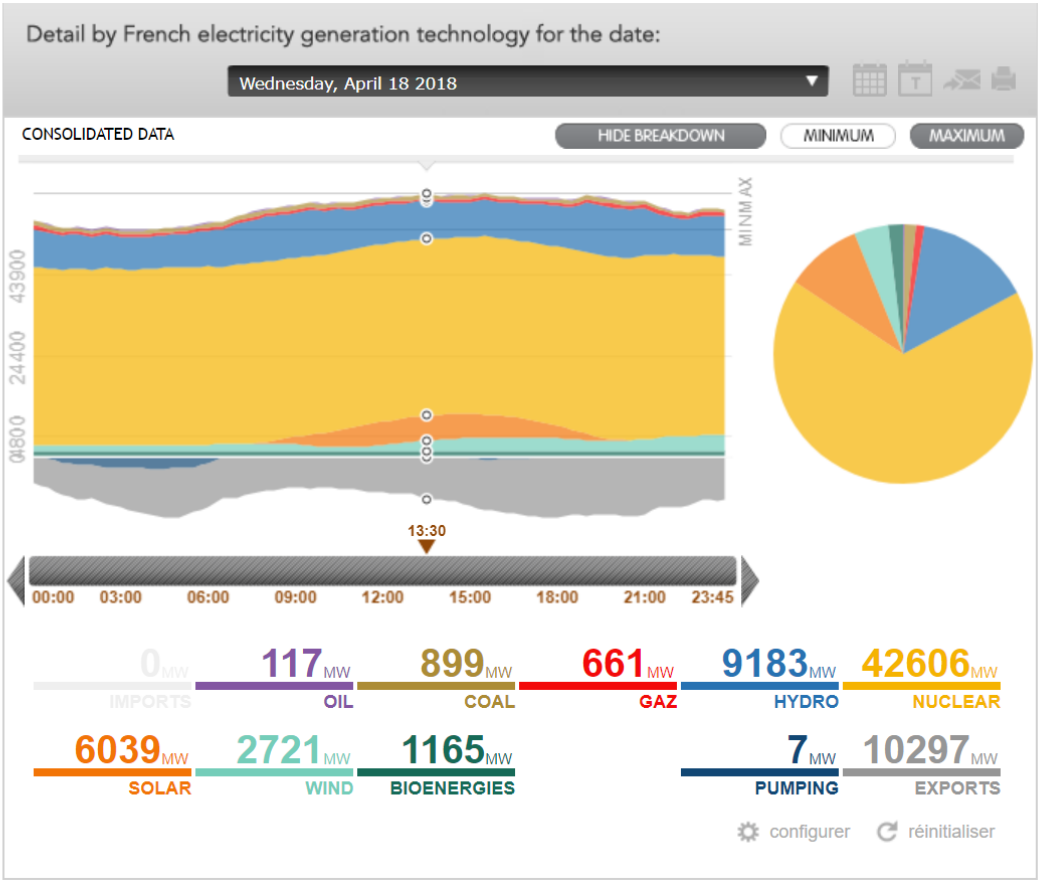
April 15th, 2019

Delaware Hackathon: Basetable Report

Introduction

For the next portion of our hackathon project, we have worked to construct an extensive basetable that will aid in our goal of predicting future electricity price in the French market. This report is designed to walk you through each component of our basetable to understand its intended purpose and get a sense of our objectives moving forward. A copy of our final basetable will be sent in a separate file alongside this report.

The prediction we are trying to do is based on how the electricity is generated based on sources. The below image captures the variation and its cause and effect relationship on the prices as electricity production cost using solar, wind is very cheap compared to sources like oil and coal.



Basetable description

Our basetable is over 35,000 observations in length (one observation per day from January 6th, 2014 until December 31st, 2018). It contains 18 variables including those concerning financial electricity market metrics, weather patterns, key calendar dates, energy generation and energy consumption. We will now explain these variables in more depth below.

Electricity market variables

Electricity market data was gathered primarily from epexspot.com, where we gathered intra-day and day-ahead information for the French market. Key focuses here were looking at hourly intra-day and day-ahead prices to compare price difference. We did not consider any values to be outliers, except for some prices which were reported to be negative (explained under IntradayPrice and DayaheadPrice._EUR_MWh variables).

fromdate

fromdate is the first variable serving to provide us the date of which all the other variables correspond to. As established, the fromdate variable covers a 4-year span.

fromtime

fromtime is the corresponding time variable (hour of day) to the fromdate variable described above. There are 24 hourly observations per day to cover every day throughout the chosen 4-year basetable period.

IntradayPrice

IntradayPrice represents the last price point reached for electricity that hour. This came directly from the data on the EPEX SPOT website. In some cases, we found the last price was indicated as a negative number. We replaced those instances with a last price of 0, as we felt it didn't make sense that a price could be negative. We also replaced all NA values by the mean Intraday price, which was €42.84. We did this to try and keep data as accurate as possible as it was a price variable, so setting them to 0 or removing them would cause further inaccuracies. We can see basic summary statistics of the intra-day Price variable below:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00   29.00   40.00   42.84   53.15  800.00
```

Dayahead_Load.Forecast

Dayahead_Load.Forecast is the forecasted electricity load required for the day-ahead market. This came directly from the data on the EPEX SPOT website. We can see basic summary statistics of Dayahead_Load.Forecast below:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
29650	45200	52400	54100	62200	95150

DayaheadPrice._EUR_MWh

DayaheadPrice._EUR_MWh is the hourly day-ahead price in euros for buying electricity. This came directly from the data on the EPEX SPOT website. In some cases, we found the last price was indicated as a negative number. We replaced those instances with a last price of 0, as we felt it didn't make sense that a price could be negative. We can see the basic summary statistics for DayaheadPrice._EUR_MWh below:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	29.99	40.12	42.57	52.07	874.01

PriceDIFF

PriceDIFF is the difference in price between the hourly last price and day-ahead price. It was computed by subtracting DayaheadPrice._EUR_MWh from IntradayPrice for each observation. Therefore, the values in this column will help us identify the difference in price and help in establishing our target variable for classification.

Target

Target is, as the name suggests, our target variable. It was largely based on the PriceDIFF variable, indicating a value of 1 if the difference was positive (price went down, therefore good time to buy) and 0 if the difference was negative (price went up, therefore bad time to buy).

Weather pattern variables

Weather data was collected for the 4-year period (2015-2018) from Paris, France. We decided to focus on one weather location as we would avoid averaging weather data with other locations and decreasing data validity. Paris was chosen due to its larger population; energy generation and consumption demand. For weather-related data, we did not consider any data to be outliers as temperature and precipitation cannot be disputed and can be unpredictable in nature. All weather data was collected via the National Climatic Data Center's historical climate database.

PRCP

PRCP indicates the amount of precipitation experienced over the course of the day. This came directly from the data via the National Climatic Data Center. All NA values were replaced with the mean precipitation value which was 0.07 millimeters per day. Basic summary statistics for this variable are as follows:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.00000 0.06805 0.06000 2.97000
```

TMAX

TMAX represents the maximum temperature (in Celsius) experienced over the course of the day. This came directly from the data via the National Climatic Data Center. All NA values were replaced with the mean maximum temperature value which was 17 degrees Celsius per day. Basic summary statistics for this variable are as follows:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.78   11.11   17.01   17.01   22.22   38.89
```

TMIN

TMIN represents the minimum temperature (in Celsius) experienced over the course of the day. This came directly from the data via the National Climatic Data Center. All NA values were replaced with the mean minimum temperature value which was just above 8 degrees Celsius per day. Basic summary statistics for this variable are as follows:

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-8.330   4.440   8.180   8.178  12.220  22.780
```

TAVG

TAVG represents the average temperature (in Celsius) experienced over the course of the day. This came directly from the data via the National Climatic Data Center. All NA values were replaced with the mean average temperature value which was around 12.4 degrees Celsius per day. Basic summary statistics for this variable are as follows:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-4.44    7.22   12.22   12.41   17.22   29.44
```

TVAR

TVAR represents the total amount of variance in temperature (in Celsius) experienced over the course of the day. This was computed by subtracting TMIN from TMAX. All NA values were replaced with the mean average temperature variance value per day which was found to be around 8.75 degrees Celsius per day. Basic summary statistics for this variable are as follows:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.550   6.110   8.750   8.754  11.110  21.120
```

TDIFF

TDIFF represents the difference in average temperature from one day to the next. This was computed by subtracting TAVG of the previous day from TAVG of the current day. For example, TDIFF for January 6th, 2015 was equal to: TAVG of January 7th - TAVG of January 6th. All NA values were replaced with 0 which was also extremely close to the mean value. Basic summary statistics for this variable are as follows:

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-7.220000 -1.120000  0.000000  0.005594  1.660000  8.330000
```

Significant

Significant indicates whether the average temperature change (TDIFF) from one day to the next was considered significant. In our case, we considered TDIFF of >= +/- 5 degrees to be a significant temperature change. If the change met this criterion, it was assigned a value of 1, otherwise it was assigned a value of 0. This is therefore a binary variable. The following table shows the ratio of significant temperature changes day over day:

```
    0     1
34032  1056
> Therefore, 3% (1056/ (1056+34032)) of days are significant
```

TPRED

TPRED represents what we estimate the temperature to be at a certain hour of each day. We used the following criteria for determining of temperature estimation:

Middle Night (01:00-04:00) = TMIN

Early Morning (05:00-08:00) = TMIN

Late Morning (09:00-12:00) = TAVG

Early Afternoon (13:00-16:00) = TMAX

Rush Hour (17:00-20:00) = TAVG

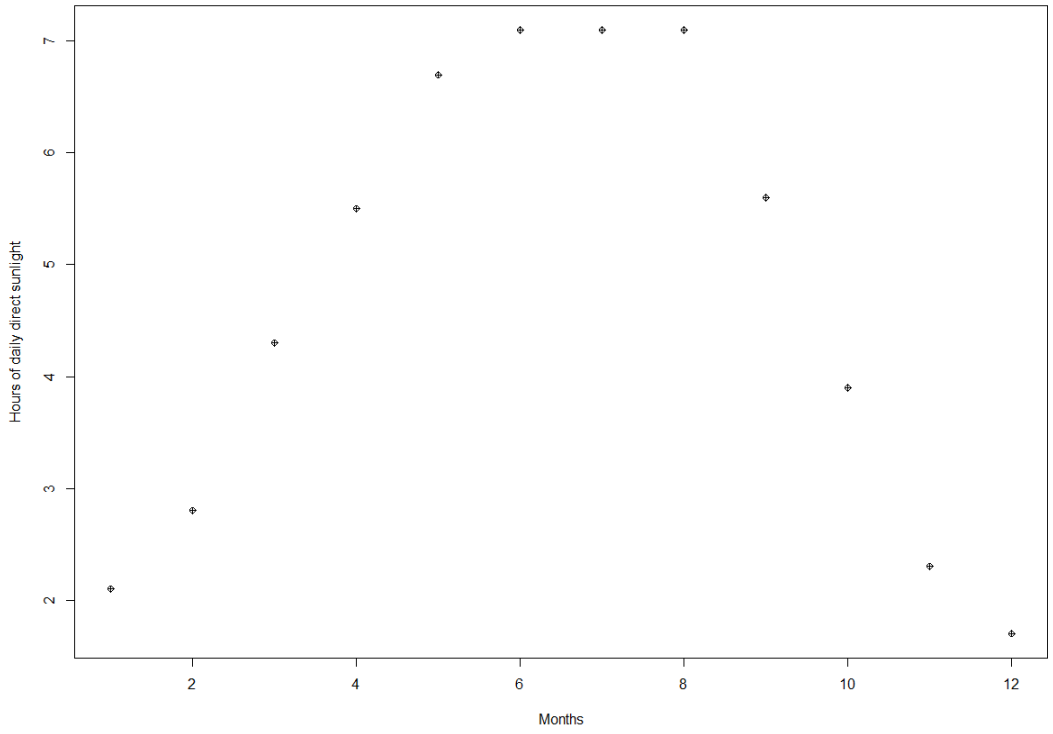
Off Peak (21:00-24:00) = TMIN

All NA values were replaced with the mean predicted temperature value per day which was found to be around 11.15 degrees Celsius per day. Basic summary statistics for this variable are as follows:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-8.33    6.11   11.15   11.15   15.56   38.89
```

DirSunHours

DirSunHours represents the number of hours of direct sunlight for each month historically in Paris. This number than was divided by the number of days in the month to achieve the average hours of direct sunlight per day. This data was collected from meteofrance.com. The following is a display of the average hours of direct sunlight per day for each month of the year:



Calendar variables

We also determined that key dates such as holidays as well as whether the day fell on a weekday or a weekend would also be an important metric to track, as obviously this would have a considerable effect on energy generation and consumption levels, as well as electricity market activity.

Weekday

Weekday is a binary variable that indicates if the day is a weekday or not. If the day is a weekday, it is assigned a value of 1, otherwise if the day falls on a weekend it is assigned a value of 0. The following table shows the ratio of weekdays to weekend days in the basetable:

0	1
24980	10108

Holiday

Holiday is a binary variable that indicates if the day is a public holiday in France or not. If the day is a public holiday, then it is assigned a value of 1, otherwise if the day is a non-holiday then it is assigned a value of 0. The following table shows the number of holiday days to non-holiday days in the basetable:

0	1
33648	1440

Next steps

Our next focus is to use this basetable to run some classification models and ultimately be able to predict whether electricity price will go up or down for the day-ahead market compared to the current intra-day ending price. We have done some elementary testing but to not have any significant results to display as of yet.

Data sources

http://www.epexspot.com/en/
http://www.meteofrance.com/climat/france
https://www.ncdc.noaa.gov/cdo-web/