

# Using multivariate polynomial regression for predicting electricity prices based on historical data

## Introduction

Since the production of electricity with renewable energy sources can't be increased as we like we will need new storage or consumption solutions for switching to renewable energy. By storing electricity when it's cheap and releasing it when it's expensive we can make renewable energy available when it's not windy or sunny. Another solution is smart consumption where appliances and electric cars are charged only when electricity is cheap. Both solutions require us to predict the price of electricity at different times.

I will create a machine learning application that predicts electricity prices in Finland based on historical data.

The price of electricity is relatively cyclical, which means you might be able to predict its price fairly accurately simply based on historical data. This is what my project is focused on. I will create a machine learning application that predicts electricity prices in Finland based on historical data.

In the first section I explain the issue as a machine learning problem and what the datapoints and its features and label are. In the second section I explain where I got the data and the models I used. Then I compare the different models and methods and choose the best in the third section. Finally I discuss the results of the chosen model and suggest improvements.

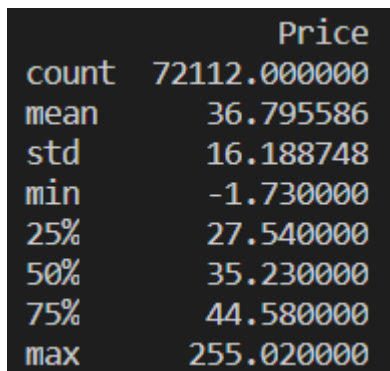
## Problem formulation

The price of electricity is relatively cyclical, which means you might be able to predict its price fairly accurately simply based on historical data. The application can then be taught how much the time of the day, day of the week, month of the year and year will affect the price of electricity and based on that predict the price in the future.

Data points in my application are one-hour timeframes. The features of the data points are **year**, **month**, **weekday** and **hour** (i.e. time of day, e.g. 14-15). The label of the data points is the price of electricity on the Finnish national electricity grid measured in euros per megawatt-hour (eur/MWh).

## Method

The dataset consists of excel files downloaded from nordpoolgroup.com. The site provides data of Finnish and other countries electricity prices every hour from the year 2013 onward, so I have 72 112 datapoints. I removed all markets except the Finnish market from the excel files. Then I modified it so that I have five columns: year, month, weekday, hour (e.g. 14 – 15), Price (in EUR/MWh). I then changed the month, weekday and hour features into integers (1-12 for months, 1-7 for weekdays and hour 0-24, e.g. 14 if time is 14-15) so that I can use a regression model for training the machine learning hypothesis.



	Price
count	72112.000000
mean	36.795586
std	16.188748
min	-1.730000
25%	27.540000
50%	35.230000
75%	44.580000
max	255.020000

*Picture 1, information about the data*

I used Python to train the weights of the models. The first method I used was multivariate polynomial regression by importing sklearn's PolynomialFeatures and transformed the features into polynomial form. Then I split the data into training, validation and test data (60% training, 20% validation, 20% test). Then I applied sklearn's LinearRegression on the training data and using it to predict the labels of the validation data. I started with a hypothesis space of a 5-degree polynomial and increased the degree to 14 and recorded the results of each hypothesis space. Sklearn's LinearRegression uses lowest Mean Squared Error (MSE) loss to optimize the weights, but I have presented the results with Root Mean Squared Error (simply the square root of MSE), since it gives a better picture of how big the errors are, but results are still the same relative to each other as using MSE. The best result came with 13-degree polynomials. Training root squared mean error was 10.6

and  $R^2$  56,3%. For the validation set the  $R^2$  was 53,0%, root mean squared error loss (RMSE) was 11.3 and mean absolute error loss (MAE) was 7.25.

There might be a difference between how the price of electricity changes during a single day or week in winter compared to a single day/week in summer. The previous method uses the same polynomial for the price changes during a day or week for the whole year. For the second method I made a small decision tree, which splits the data into four seasons and then uses multivariate polynomial regression for the remaining features. This way I have used the decision tree model and four different regression models, one for each season. The regression part was done the same as previously except it was applied four times and used one feature less (month was not taken into account). The best results came again with 13-degree polynomials. The results for the training error were: RMSE: 12.2, MAE: 8.05, for the validation set: RMSE: 12.7, MAE: 8.26.

## Results

Here are the recorded results from the models (with the best result highlighted in green):

Method1	Polynomial degree	Training error:			Validation error:		
		RMSE	MAE	$R^2$	RMSE	MAE	$R^2$
	5	12.23225	8.123237	0.421531	12.72438	8.266788	0.408687
	6	12.01714	8.088724	0.441697	12.52174	8.286633	0.427371
	7	11.68033	7.660633	0.472554	12.19807	7.876562	0.456591
	8	11.16325	7.289828	0.518219	11.70097	7.506577	0.499979
	9	11.01135	7.145412	0.531241	11.55861	7.378305	0.512072
	10	10.9478	7.126755	0.536636	11.55155	7.397474	0.512668
	11	10.91632	7.168084	0.539298	11.53435	7.476265	0.514118
	12	10.84112	7.073657	0.545623	11.48362	7.374735	0.518382
	13	10.62585	6.896461	0.563489	11.34202	7.251208	0.530187
	14	10.63226	6.924547	0.562962	11.3813	7.303322	0.526927

Method2	Polynomial degree	Training error:		Validation error:	
		RMSE	MAE	RMSE	MAE
	8	12.2867	8.158191	12.77108	8.348743
	9	12.2261	8.082989	12.72095	8.283193
	10	12.29362	8.106337	12.79436	8.306041
	11	13.83999	9.70904	14.28534	9.894172
	12	13.22827	9.135182	13.63093	9.270229
	13	12.17295	8.049101	12.67917	8.263595
	14	12.28653	8.288298	12.81456	8.53112

The training, validation and test errors were all very similar in both methods used, which means the hypothesis spaces chosen were good. There is little chance of overfitting anyway, since I had over 72 thousand datapoints. I included in the results RMSE (Root Mean Squared Error), but I also presented MAE. MAE was included, since RMSE gives a high weight to large errors, which is not necessarily desirable.  $R^2$  was also included in the first method since it was easy to extract with the python functions that I used and  $R^2$  is also very useful, since it tells how much of the price can be explained by the chosen features.

The first method of only using multivariate polynomial regression was better than splitting into four classes and then using polynomial regression. Best result of the first method test errors: RMSE = 10.8, MAE = 7.16, second method: RMSE: 12.2, MAE: 8.19. So the 13-degree polynomial hypothesis space for the first method is chosen for this application. I then calculated the error values for the **test set** (which has not been used to train the model or to choose between hypothesis spaces):  **$R^2 = 55,3\%$ , RMSE = 10.8, MAE = 7.16.**

## Conclusion

The chosen model has a test set  $R^2$  of 55,3%, meaning only about half of the price variation can be explained by what date and time it is. This isn't really enough.

To improve the model, you could add more features, e.g. weather conditions in Finland or analyze electricity production in Finland (how much is renewable, nuclear, fossil and production cost and variation of these sources).

## References

<https://www.nordpoolgroup.com/historical-market-data/>