Predicting Electricity Consumption Using Weather Data

Sejin Jang · Follow
5 min read · Sep 15, 2021

Listen Share More

This article showcases a machine learning project that uses multivariate time series and Vector Auto Regression (VAR) model.

Photo courtesy of Mark Lennihan/AP

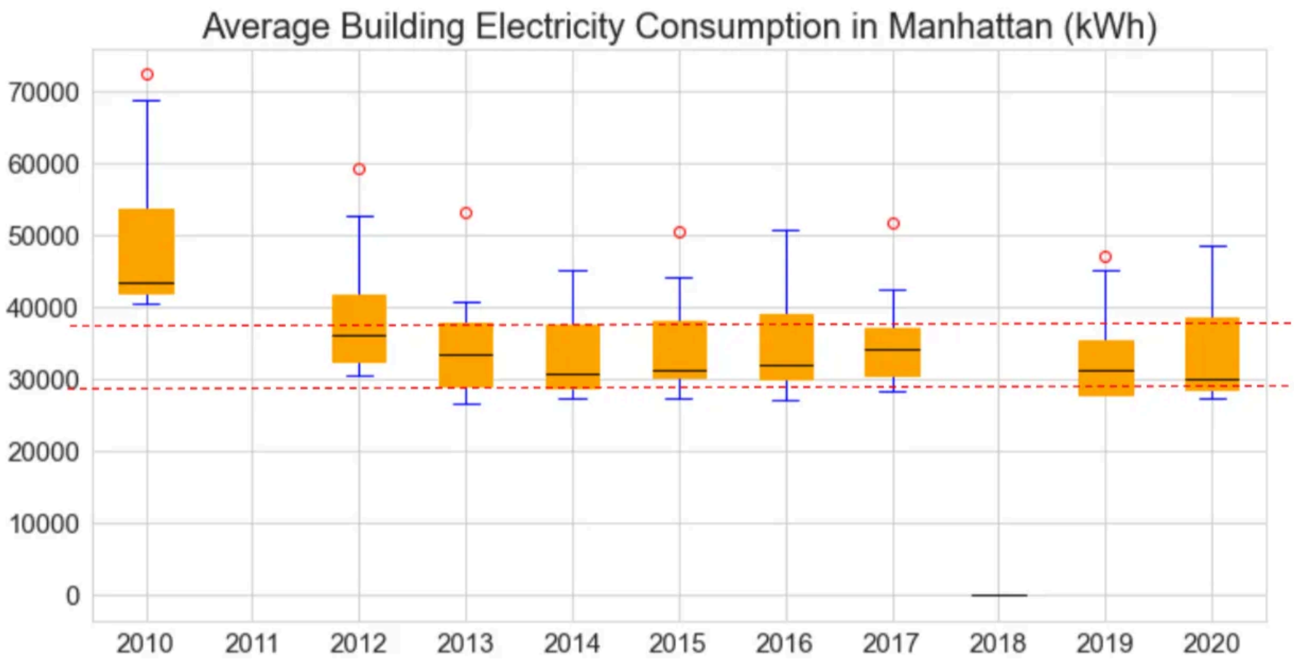
According to the UN climate report in 2021, some of the global warming effects are now irreve … inue to rise, due to … el are at high risk. Furthermore, as the temperature increases, this also results in extreme weather events such as hurricanes, wildfires, and drought. To mitigate global warming, some have already resorted to using clean and renewable energy sources. This can be seen as more communities are trying to replace their traditional energy sources that rely on fossil fuels, with clean energy such as solar and wind. Hence, it is crucial to estimate the future energy demand so that a more clear and realistic goal can be set for the communities.

In this article, I will share my machine learning project that created a predictive model estimating the future electricity consumption in New York City with its weather data. Energy usage is correlated with the temperatures. People are more likely to use more air conditioning in hot weather and more heaters during cold periods.

Two datasets were used in this project — one on the electricity consumption and the other on weather. The consumption dataset was found at the NYC Open Data website. It contains monthly electricity consumption and cost data from January 2010 to February 2021 by borough and development. It includes features such as electricity consumption, utility vendor and meter information. The electricity consumption data is provided in kilowatt (kW) as well as in kilowatt-hours (kWh). The difference between kWh and kW is that kW reflects the rate of electricity you use, and kWh indicates the amount of electricity you use. For this project, kWh is taken into account.

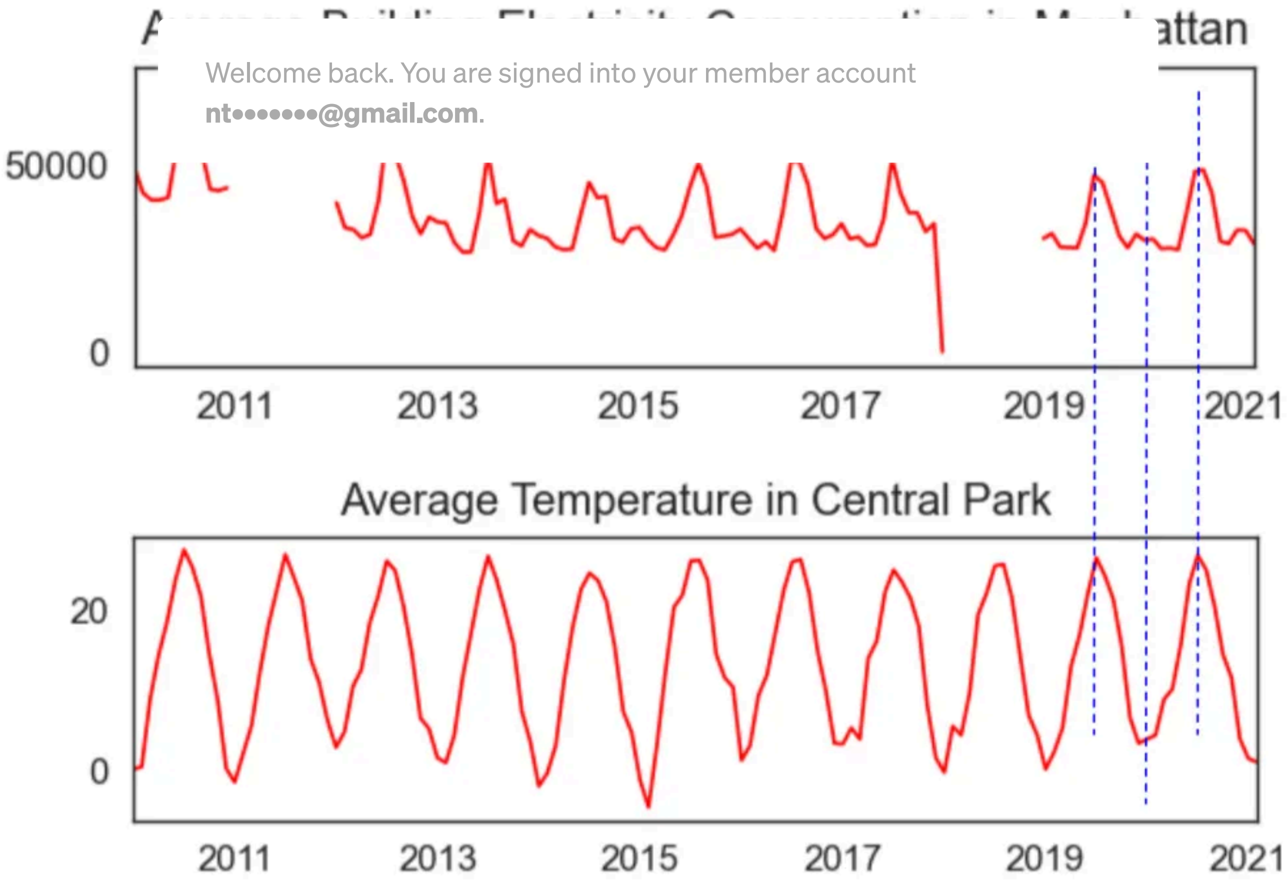
The weather dataset was gathered from the National Oceanic and Atmospheric Administration (NOAA), an American scientific and regulatory agency within the United States Department of Commerce. It contains monthly mean maximum, mean minimum and mean temperatures; monthly heating and cooling degree days; and extreme daily temperature.

The NYC electricity consumption dataset contains 362,630 records. Each record represents a building and a month. It has data from 195 buildings in Manhattan, 69 buildings in the Bronx, 262 buildings in Brooklyn, 93 buildings in Queens and 10 buildings in Staten Island. This is obviously not all the data in New York City as the number of buildings in the dataset is small. Thus, I decided to focus on Manhattan as a sample



The graph above is showing a building's electricity consumption in Manhattan on average. Except for 2010, which seems rather higher than the other years, and 2011 and 2018 with incomplete and missing data, the Median electricity usage in a year ranges from around 30,000 to 37,000 kWh.

To match the location, I took Central Park's different types of temperatures in celsius (°C). After merging the two datasets, I could see that weather as well as the electricity consumption have seasonality. The highest consumption is observed in the Summer. The next highest consumption is observed in the Winter.



Now, to predict the electricity consumption in any given month, I used models such as Vector Auto Regression (VAR). The basis behind these models is that this month's electricity consumption can be affected not only by the last month's consumption but also by the last month's temperature. This means that you can predict the future electricity consumption with past values of itself along with past values of temperature.

Before fitting the VAR model, all the time series are made stationary by using a differencing technique. The key for VAR modeling is that you have to find the optimal order (lag) value. To find this, you can use an attribute called **.select_order()**.

```
model.select_order()
```

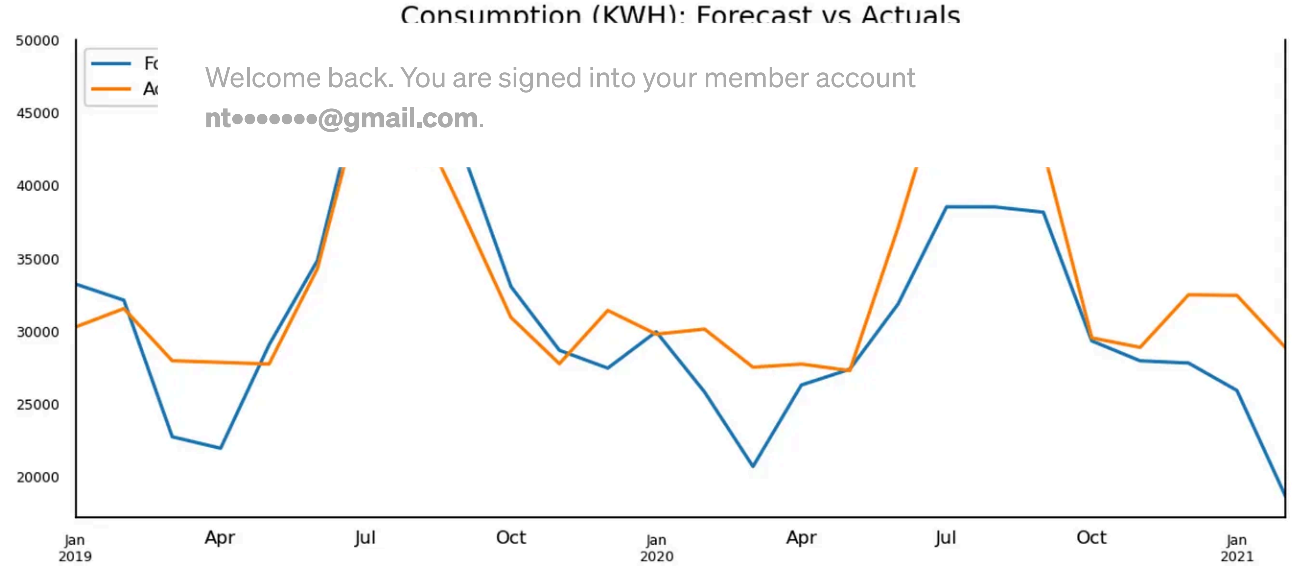
	AIC	BIC	FPE	HQIC
0	47.65	47.95	4.938e+20	47.76
1	44.80	47.48	2.950e+19	45.83
2	42.39	47.44	3.090e+18	44.33
3	40.60	48.03	8.169e+17	43.45
4	37.90	47.71	1.585e+17	41.67
5	23.93*	36.13*	1.481e+12*	28.62*

After fitting the model with the optimal order, 5 in this case, I get the equations for all the time series variables. Here, I only care about the consumption time series so I focus on the equation for the consumption as seen in the image below. The way I interpret this is that first, I observe which variable is most influential in prediction by looking at the p-values. I consider those under 0.05 are the influential variables, so in this case, those in red boxes are contributing to the consumption prediction the most. My equation is made up of coefficients of all the time series.

```
Results for equation Consumption (KWH)
============                                                                          =========
                                                                                           prob
------------                                                                          ---------
const                                                                                     0.797
L1.Consumpti                                                                              0.000
L1.Cooling L                                                                              0.878
L1.Extreme Minimum Temperature      -1221.702665      332.085650          -3.679          0.000
L1.Extreme Maximum Temperature       -935.981988      412.752088          -2.268          0.023
L1.Heating Degree Days Season           5.158869        1.603314           3.218          0.001
L1.Average Temperature               3036.170960    13121.979882           0.231          0.817
L1.Maximum Temperature               -826.095799     6611.897309          -0.125          0.901
L1.Minimum Temperature               -461.802277     6450.405015          -0.072          0.943
L2.Consumption (KWH)                   -0.726849        0.158762          -4.578          0.000
L2.Cooling Degree Days Season           1.705921        5.540067           0.308          0.758
L2.Extreme Minimum Temperature      -1153.012777      480.257721          -2.401          0.016
L2.Extreme Maximum Temperature      -1000.420098      619.066632          -1.616          0.106
L2.Heating Degree Days Season           2.617685        2.043876           1.281          0.200
L2.Average Temperature              39624.410938    21365.877955           1.855          0.064
L2.Maximum Temperature             -20690.393893    10534.495839          -1.964          0.050
L2.Minimum Temperature             -16844.698886    10615.841066          -1.587          0.113
L3.Consumption (KWH)                   -0.935593        0.197398          -4.740          0.000
L3.Cooling Degree Days Season          -2.131049        5.450751          -0.391          0.696
L3.Extreme Minimum Temperature       -286.397127      442.370556          -0.647          0.517
L3.Extreme Maximum Temperature        -88.478301      561.751384          -0.158          0.875
L3.Heating Degree Days Season           3.915955        2.143510           1.827          0.068
L3.Average Temperature              43769.437810    20360.854436           2.150          0.032
L3.Maximum Temperature             -20600.285732    10099.218114          -2.040          0.041
L3.Minimum Temperature             -21854.841476    10193.842960          -2.144          0.032
L4.Consumption (KWH)                   -0.321195        0.195149          -1.646          0.100
L4.Cooling Degree Days Season           0.190374        6.119921           0.031          0.975
L4.Extreme Minimum Temperature        885.885277      387.111808           2.288          0.022
L4.Extreme Maximum Temperature        976.781726      440.302316           2.218          0.027
L4.Heating Degree Days Season           8.086861        2.167904           3.730          0.000
L4.Average Temperature              11198.927922    16409.561084           0.682          0.495
L4.Maximum Temperature              -7016.570081     8112.943203          -0.865          0.387
L4.Minimum Temperature              -5528.204644     8270.904203          -0.668          0.504
L5.Consumption (KWH)                    0.031650        0.148366           0.213          0.831
L5.Cooling Degree Days Season         -17.816074        6.512404          -2.736          0.006
L5.Extreme Minimum Temperature        832.889337      253.365486           3.287          0.001
L5.Extreme Maximum Temperature        861.130615      271.385606           3.173          0.002
L5.Heating Degree Days Season           6.821983        1.436667           4.748          0.000
L5.Average Temperature              34802.760640    10673.428132           3.261          0.001
L5.Maximum Temperature             -18653.658265     5390.678909          -3.460          0.001
L5.Minimum Temperature             -16406.300035     5184.456909          -3.165          0.002
===============================================================================================
```


The models are evaluated with RMSE (Root Mean Squared Error), which shows how much kWh are off, and MAPE (Mean Average Percentage Error), which expresses the forecast error by how many percentage points the forecasts are off on average.

My best model's prediction is off by 4,800 kWh, which indicates about 11% of



In conclusion, the weather data is useful for predicting energy demand. Also, more data will certainly improve the model performance.

I hope that this project inspires more researchers to try to estimate the future energy demand using weather data and help decision makers to better prepare for the demand while adopting more clean and renewable energy.

Check out the full analysis at https://github.com/cezine/ElectricityDemand

Climate Change | Energy | Data Science | Machine Learning | New York City

Follow

Written by Sejin Jang

9 Followers

Aspiring Digital Nomad — Interested in Data Science. Project Management. International Affairs, Music, Nature and Fo

More from Sejin Jang

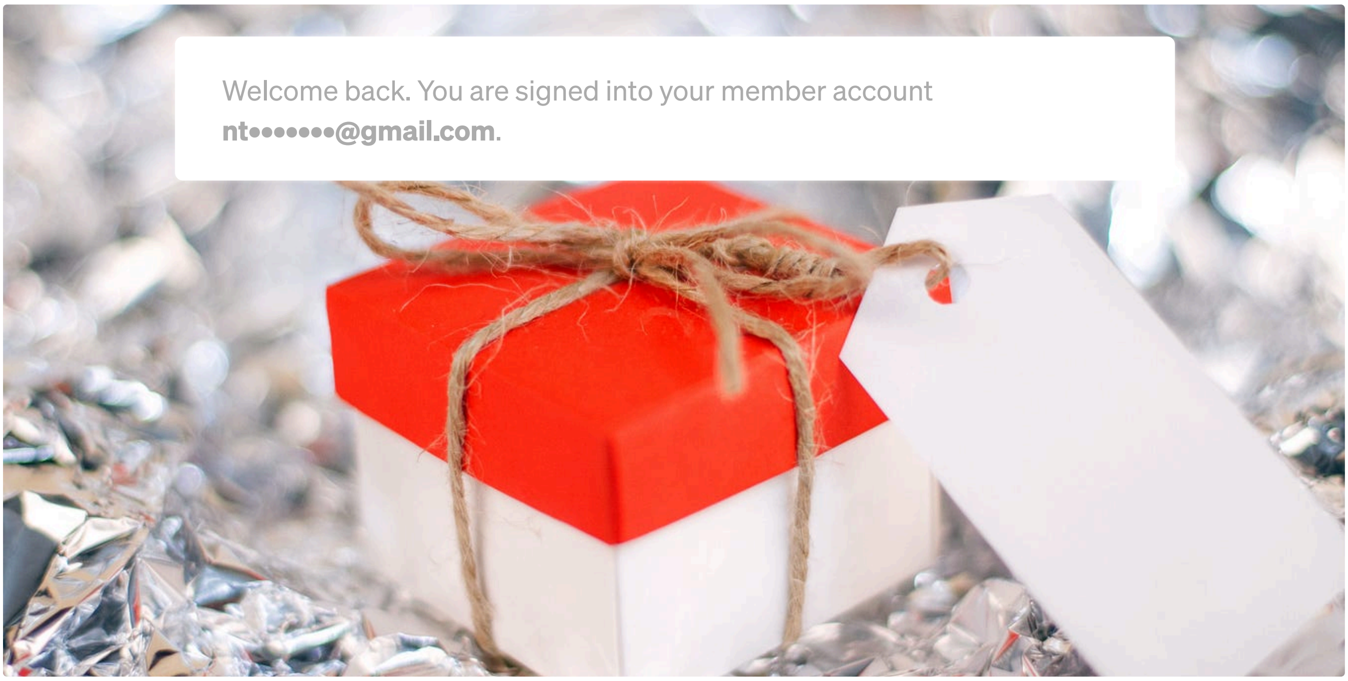
 Sejin Jang

Introduction to Machine Learning Pipelines

A machine learning workflow starts with preparing a dataset. You would explore the data, clean it, transform it, and run a machine learning…

· 3 min read · Aug 11, 2021

9



Joshua Phuong Le

Guide to Python Project Structure and Packaging

TIPS:

10 min read · Feb 3, 2023

341 4

Maximilian Vogel

The 10 Best Free Prompt Engineering Courses & Resources for ChatGPT, Midjourney & Co.

Updated Jar

8 min read · S

1.2K 16

Sejin Jang

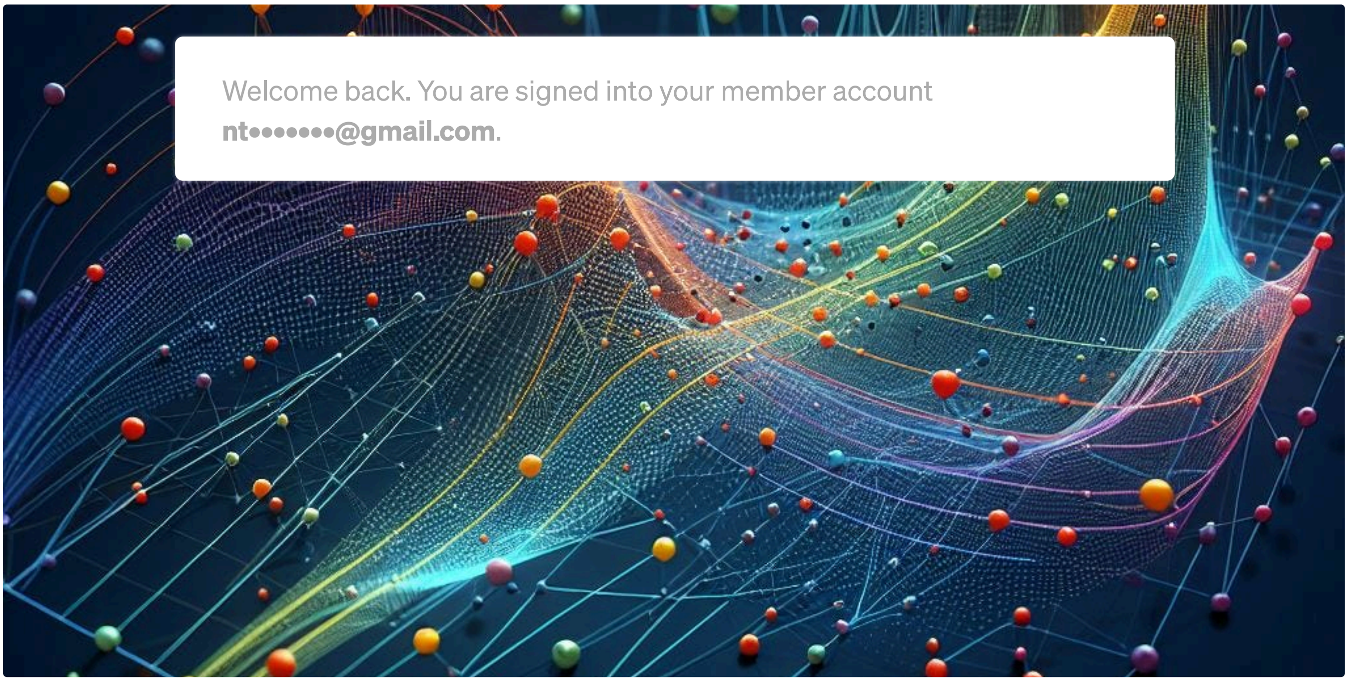
Moneyfilm — A fun Exploratory Data Analysis on Movies

Exploratory data analysis (EDA) is a popular approach to analyze data sets and summarize their characteristics. It can help discover...

4 min read · Jul 20, 2021

See all from Sejin Jang

Recommended from Medium



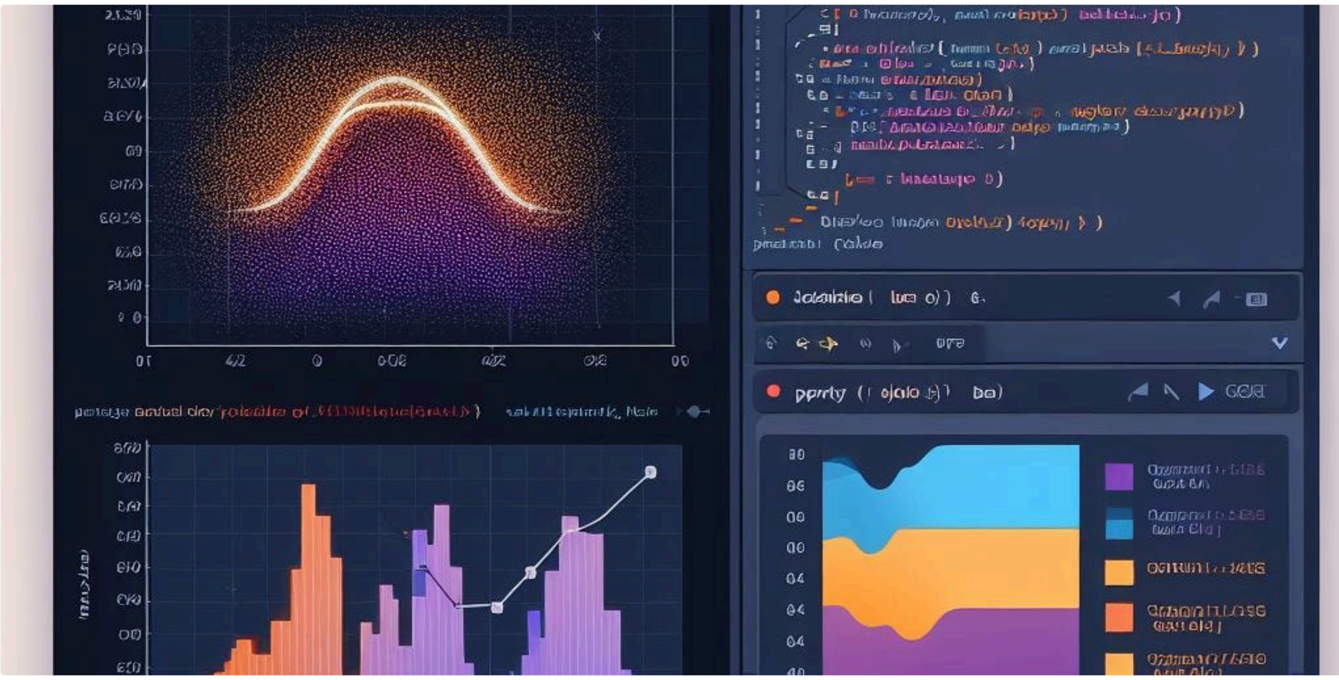
Tim Sumner in Towards Data Science

A New Coefficient of Correlation

What if you were told there exists a new way to measure the relationship between two variables just like correlation except possibly...

10 min read · Apr 1, 2024

3K 36



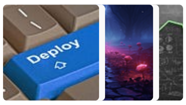
Daniel Wu

Elevate Your Python Data Visualization Skills: A Deep Dive into Advanced Plotly Tec

Data visualiz ing insights into the underlying patterns, trends...

8 min read · Nov 22, 2023

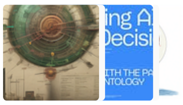
Lists



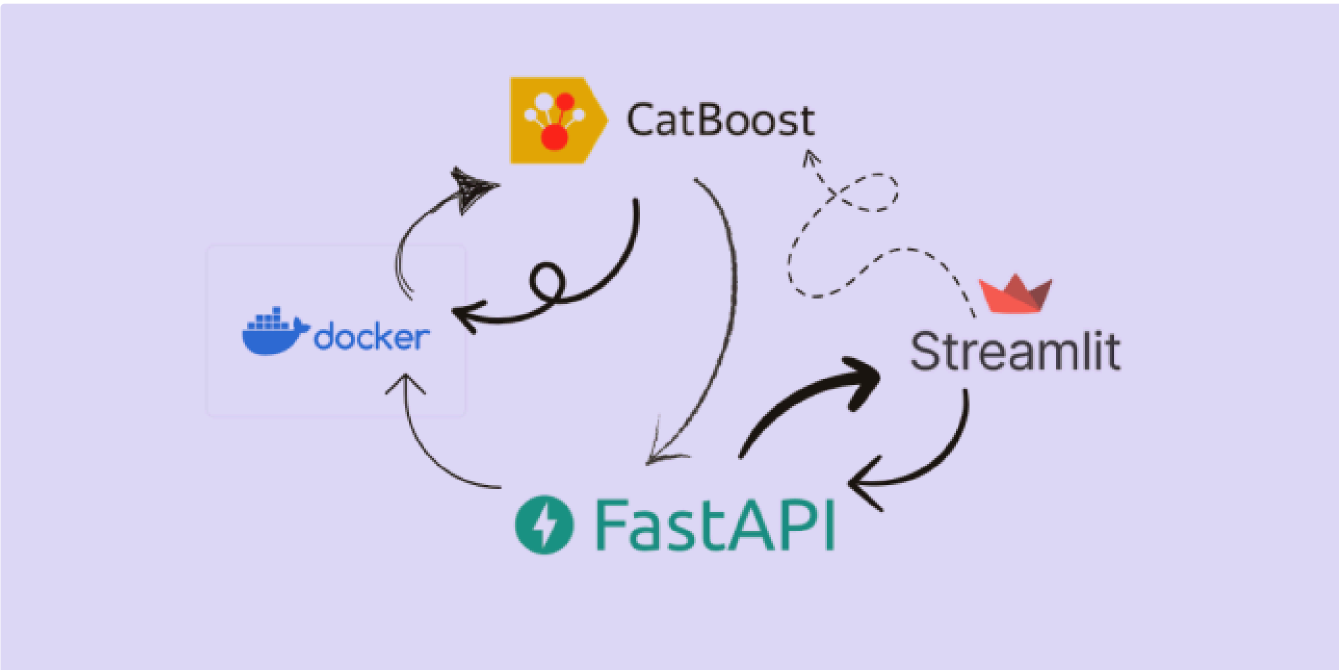
Predictive Modeling w/ Python
20 stories · 1125 saves

Practical Guides to Machine Learning
10 stories · 1349 saves

Natural Language Processing
1399 stories · 898 saves



data science and AI
40 stories · 137 saves



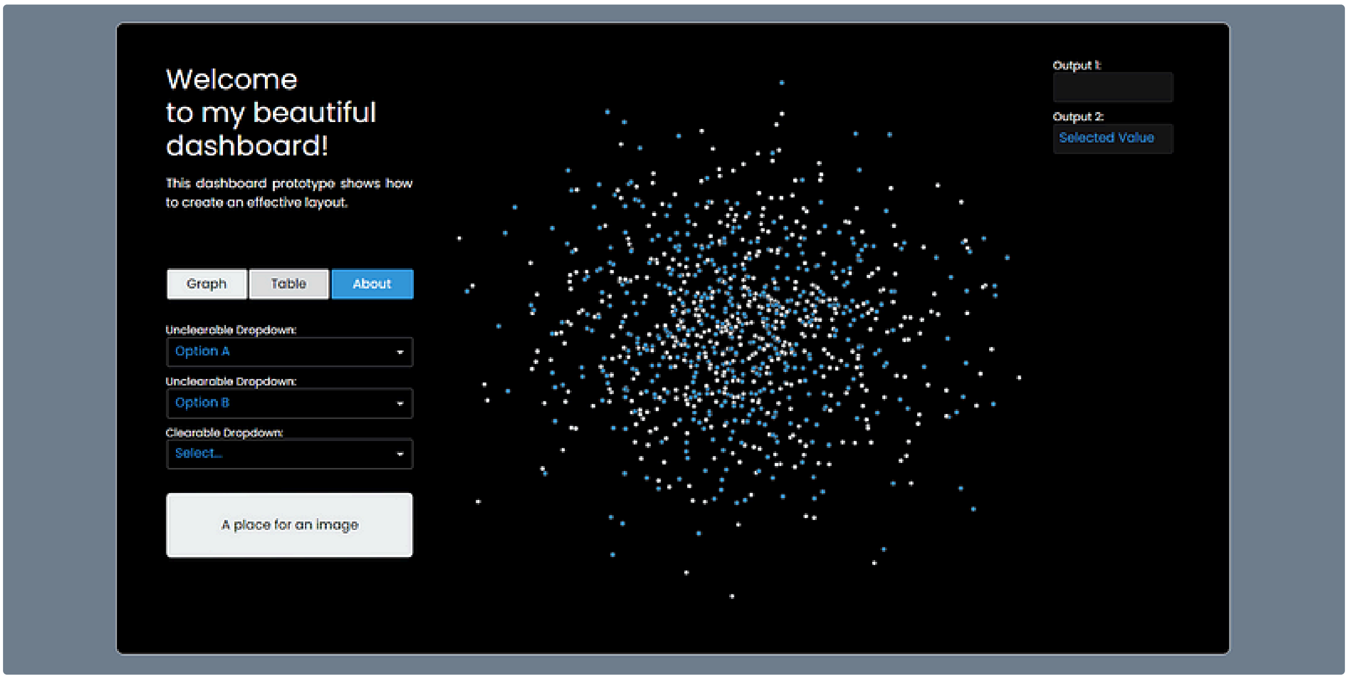
Ramazan Olmez

End-to-End Machine Learning Project: Churn Prediction

The main objective of this article is to develop an end-to-end machine learning project. For a model to be truly useful, it needs to be...

18 min read ·

87



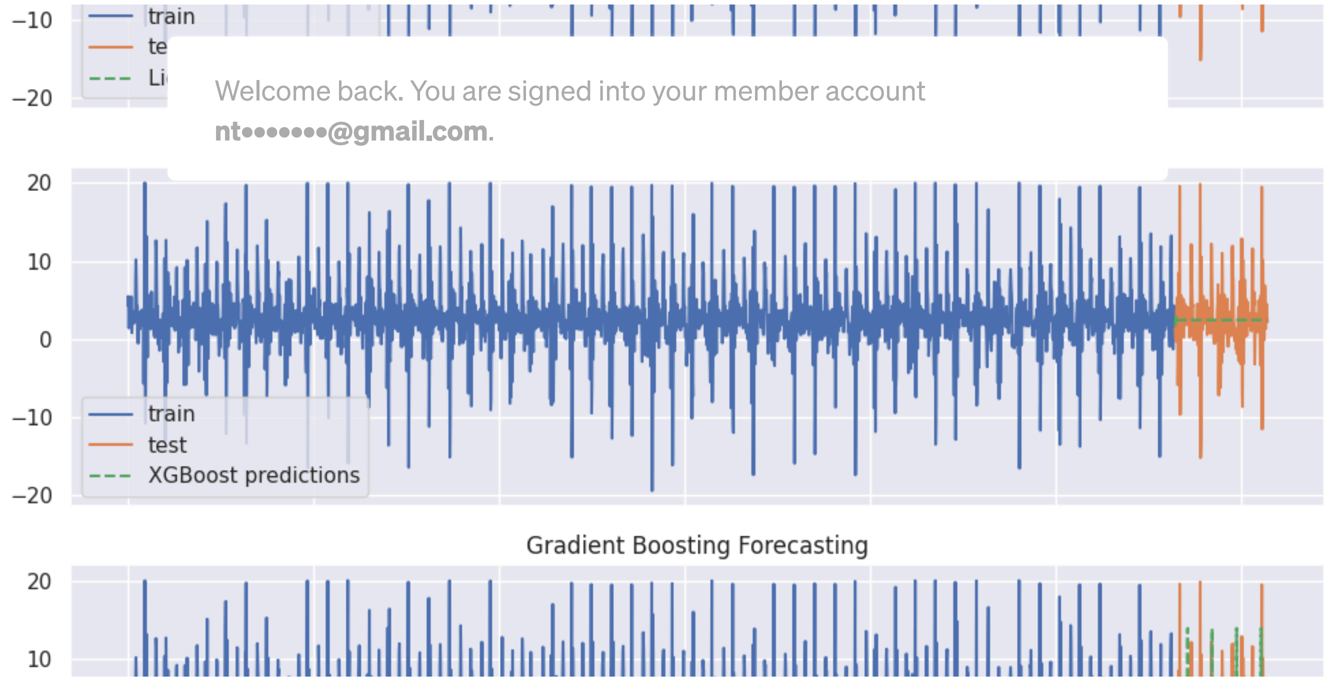
Tanya Lomskaya in Plotly

How to create a beautiful, interactive dashboard layout in Python with Plotly Dash

When created in Python, a dashboard can have an impressive design, unique interactivity, and the highest processing speed. It sounds great...

19 min read · 6 days ago

153



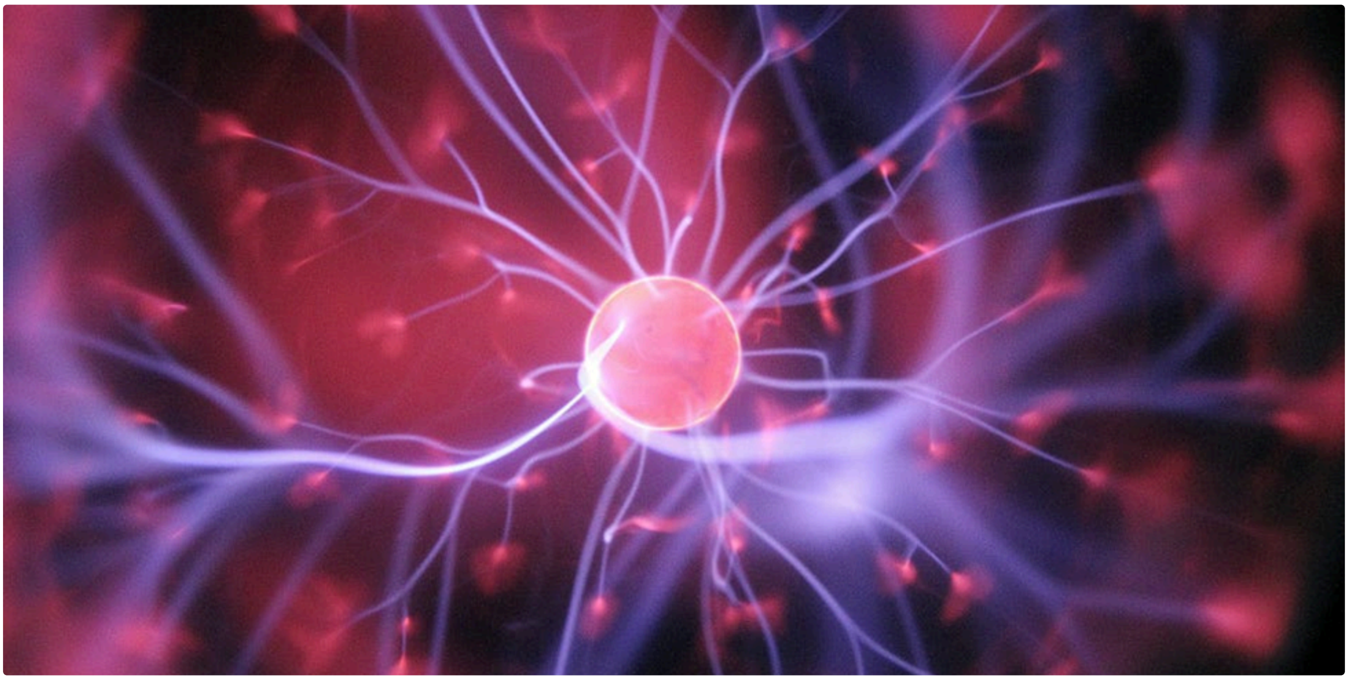
daython3

Master Time Series Forecasting in Python with skforecast

A Guide to the Python Library for Time Series Forecasting

· 5 min read · Nov 24, 2023

143



Shenggang Li in DataDrivenInvestor

Time Series Forecasting: A Comparative Analysis of SARIMAX, RNN, LSTM, Prophet, and Transformer...

Assessing th Efficiency and Efficacy of Leading Forecasting Algorithms Across Diverse Datasets

· 12 min re

100 1

See more recommendations