AI-Driven Exploration and Prediction of Company Registration Trends With Registar Of Companies(RoC)

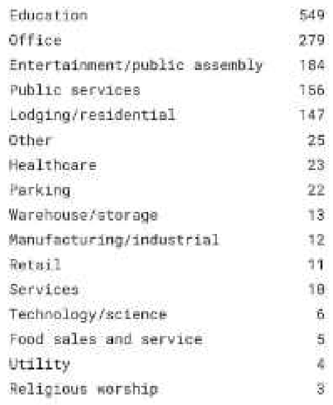
Submitted By: Velluru Vennela

Register No: 723921243053

Time Series Forecasting: Ensemble Learning

In this article, I am going to showcase the predictive power of ensemble learning for time series forecasting. Ensemble learning leads to models with higher predictive accuracy, a lower likelihood of overfitting, and a diverse set of predictions.

I will be using a dataset provided by ASHRAE (The American Society of Heating, Refrigerating and Air-Conditioning Engineers) which has hourly meterage data on electricity, chilled water, steam, and hot water for various buildings. This dataset can be found here (CC0: Public Domain). In the following section, we will import the necessary python packages, and load the meterage and building data.



```
Education                        549
Office                           279
Entertainment/public assembly    184
Public services                  156
Lodging/residential              147
Other                             25
Healthcare                        23
Parking                           22
Warehouse/storage                 13
Manufacturing/industrial          12
Retail                            11
Services                          10
Technology/science                 6
Food sales and service             5
Utility                            4
Religious worship                  3
```

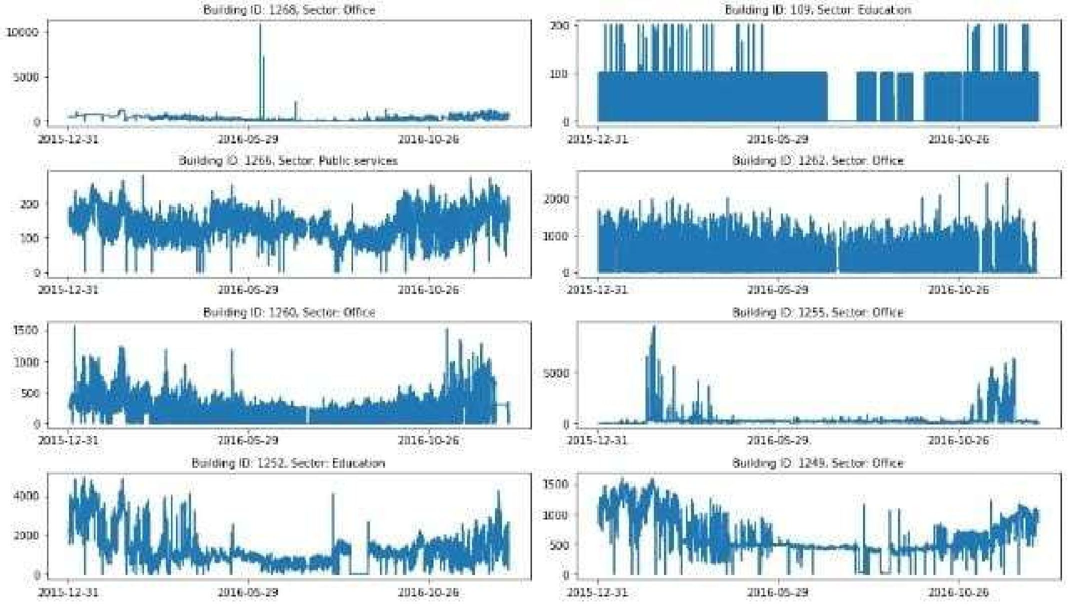
Building Sectors — By Author

In the visualization above, we can see that there are ~1400 unique buildings across numerous sectors. Not all of the buildings have clean data, with meterage data from some buildings containing more missing values than others. Let’s visualize a sample of buildings with the least number of missing data points.

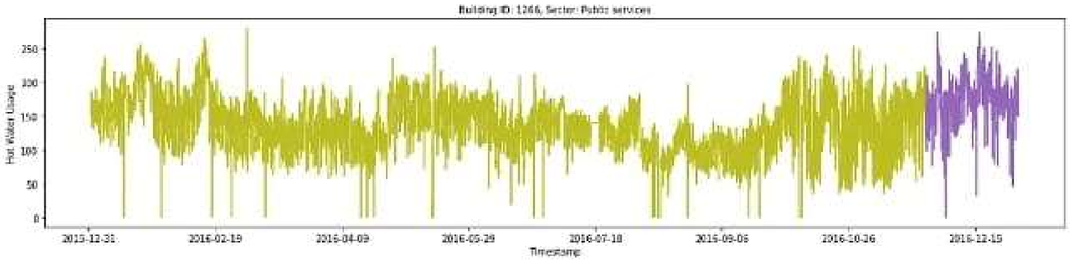
We can see that most of the plots are quite different from one another. For example, even for the office buildings, we can see that the water usage across each building is quite different. This suggests that the sector of a building may not be very useful for each prediction.

For the remainder of the article, we will be working with the hot water usage data for building 1266. This seems to have few errors or long periods of no usage. Using various features and external datasets we will hopefully create a model with high predictive accuracy. In the following cell, we isolate this building sample, split the sample into a training and validation set, and plot the result. The olive data is the training data, and the purple section is the validation set. It appears that there are 8–10 spikes in hot water usage over the course of the year. These could be due to low temperatures, special events, or something else entirely. There are also 10–20 points of 0 hot water usage for the building which could indicate holidays, weekends, or errors. These preliminary ideas can not be confirmed or denied using visualizations alone, and we will need to perform further exploration when developing the model.

We will now look at potential seasonality using time-based features from the data. We can create subplots in Matplotlib, and take the average metered hot water usage for each possible feature value.

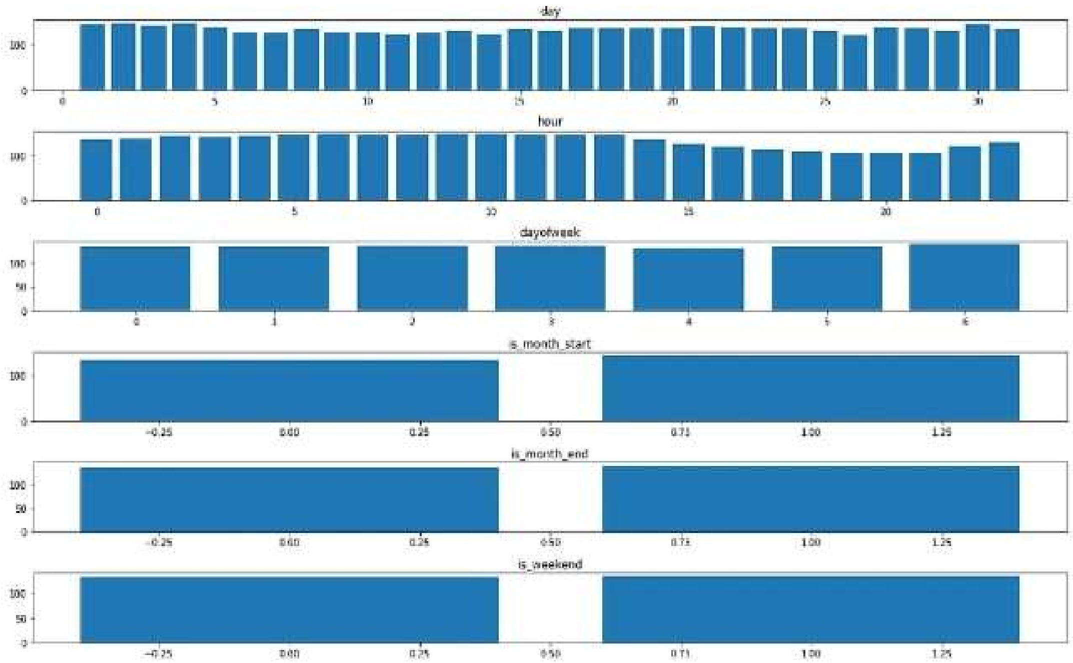


Building Samples — By Author



Building 1266 Hot Water Usage — By Author

There seems to be more hot water usage during morning hours (AM) versus the afternoon hours (PM). Interestingly, there also seems to be higher hot water usage on Sunday than on any other day in the week. The day, day of the week, and other features may be important or they could be just noise, but we can not determine this from visualizations alone.

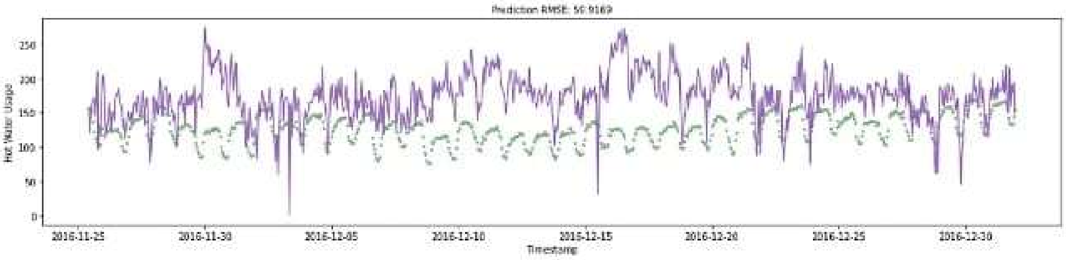


Potential Seasonality — By Author

Baseline LGBM Model

By using these features alone we can generate a baseline LGBM model. LGBM is a gradient boosting framework based on the decision tree algorithm. This framework differs from XGB as its tree models grow leaf-wise rather than depth-wise. This results in faster training times and comparable accuracy.

We can then use the trained model to predict the hot water usage for every hour in the validation set and plot the predictions with the true labels. The plot title also shows the RMSE (root mean squared error) for the predictions. This will give us a concrete metric that will use to measure model accuracy. Note that we are not yet doing an n-step ahead prediction, and we are purely using features engineered from the timestamp feature.



Baseline LGBM RMSE: ~59.9 — By Author

Holiday Data

In the next section, we will be using an external data source of US holiday data (CC0: Public Domain). This could be a useful feature in explaining the troughs in hot water demand and could potentially improve the model accuracy. After merging this data into the training set, we will update the model and make new predictions.

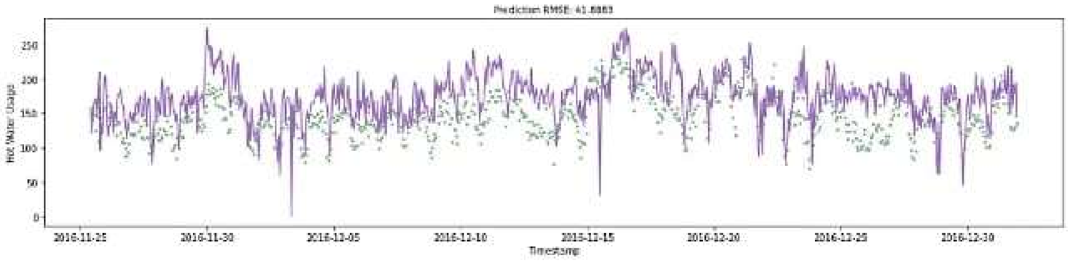
This model improved on the previous RMSE by 0.0041. This is quite an underwhelming improvement, but we will keep this feature for now in case it is valuable in combination with other features we add in the next section.

Weather Data

The second data source that I will add is weather data from the ASHRAE dataset (CC0: Public Domain). This provides real-time historical weather measurements from towers closest to the buildings in the dataset.

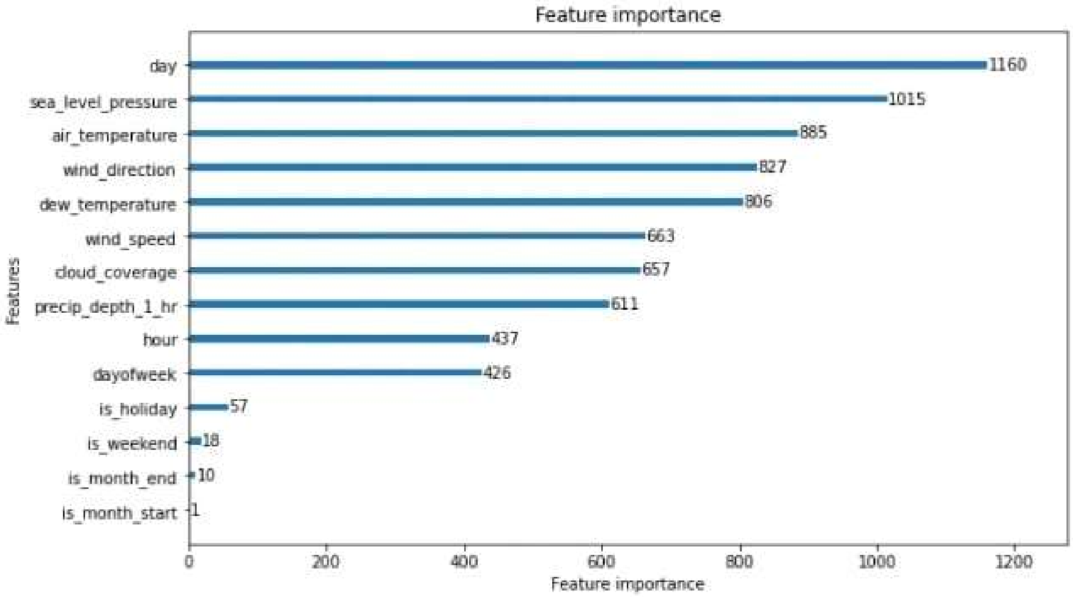
We have to be careful here as the data we have is from real-time measurements and not a 10-hour ahead forecast. If we simply used this weather information as a feature we would be committing data leakage! Therefore we will simulate error in the weather data by adding a small amount of Gaussian noise to each column. This should make the model less dependent on the weather data and less prone to overfitting.

We can then fit the LGBM model with the added features, score the RMSE of the new predictions, and plot the predictions against their true values.



LGBM w/ Weather RMSE: ~41.8 — By Author

Wow! That’s a significant increase of ~20 points on the RMSE from the previous iteration. Since we added 7 features to this iteration of the model it’s not easy to see which contributed the most to the improved accuracy. Thankfully LGBM models have feature importance built into the models themselves. In the following plot, we visualize each feature’s importance. We can see that all 7 features are important to the model, but the sea level pressure, air temperature, and wind direction have the most significant effect.



Feature Importance Plot — By Author

ARIMA Model

Finally, we will incorporate an ARIMA model in our predictions. In the real world, it does not make sense to make a prediction for the next 750 hours at a single point in time. It is more likely that you would forecast one data point n-hours ahead of time. The further ahead the prediction you make, the less accurate it is likely to be.

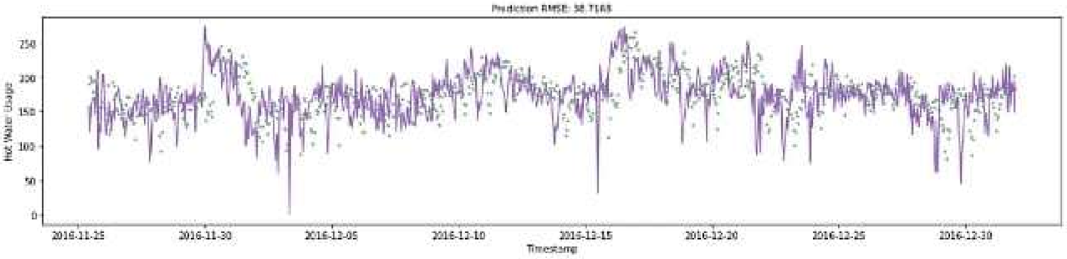
One of the benefits of doing an n-step ahead prediction is that you can incorporate more localized information into your model. We will do this by ensembling an LGBM model with a SARIMA (seasonal autoregressive integrated moving average) model. SARIMA is based on the ARIMA model architecture, but it incorporates a seasonal component. ARIMA-based models are the gold standard of time series forecasting and yield accurate predictions.

In the next cell, we are iteratively making 10-step ahead predictions on the entire validation dataset and then seeing how well the SARIMA performs on its own.

The gold standard of time series forecasting strikes again! The model achieves an RMSE of ~38 which outperforms the LGBM model by almost 3 points.

One might think that dropping the LGBM model is the way to go, but we can likely achieve even better model accuracy by ensembling the two models. Ensembling is when you take the average predictions of multiple models to make a final prediction. This often increases model accuracy and reduces overfitting.

Wow! We improved the RMSE of the SARIMA model by ~7 points and of the LGBM model by ~9 points. This is a significant improvement and showcases the power of ensemble learning.



SARIMA model RMSE: ~38.7 — By Author

Final Thoughts

Ensemble learning is just as powerful in AI/ML as in the real world. In the real world when you solve a problem with a diverse group of people you are likely to create more effective solutions. This is due to each group member's unique experience and diverse skill sets. This is no different from ensemble learning. By combining different model architectures and feature combinations you are creating a diverse set of predictions.

If you want further information I have included some additional resources below. The code for this article can be found here.

Resources

1.Ensemble Learning Wiki

2.Time Series Forecasting w/ ARIMA, SARIMA and SARIMAX

3.LGBM Documentation

4.StatsModels Documentation