A close up of text on a white background

Description generated with high confidence

**Contents**

[Introduction 3](#_Toc494457009)

[Demo Use Case Overview 3](#_Toc494457010)

[Dataset 4](#_Toc494457011)

[Model Used 4](#_Toc494457012)

[Prerequisites 5](#_Toc494457013)

[Demo components 5](#_Toc494457014)

[Data Preparation 5](#_Toc494457015)

[Model Training 7](#_Toc494457016)

[Model Application 7](#_Toc494457017)

[Model Update 8](#_Toc494457018)

[Model Score 9](#_Toc494457019)

[Dashboards 10](#_Toc494457020)

[Demo Summary 10](#_Toc494457021)

[Further Development 11](#_Toc494457022)

[Additional Reading 11](#_Toc494457023)

[Contact Details 12](#_Toc494457024)

# Introduction

This demo was created based on London Air Quality demo developed by Mark Semenenko ([msemenenko@pentaho.com](mailto:msemenenko@pentaho.com)). Please contact Mark to get access to the demo environment (AWS instance).

## Demo Use Case Overview

Air pollution causes 40,000 early deaths a year in the UK. 9,000 of those affected are residents of London. The health problems resulting from exposure to air pollution also have a high cost to society and business, our health services, and people who suffer from illness and premature death. In the UK, these costs add up to more than £20 billion every year. In addition to these costs the EU Commission has enacted legislation that cities in the EU must abide by or risk being fined.



Figure 1. Photo from Great London Smog of 1952 (www.history.com)

Prediction and monitoring of air quality in big cities like London is crucial for city councils. It allows informing citizens in timely manner and send notifications. In this way, residence can protect themselves and others by avoiding the car and using back streets to walk and cycle instead.

This demo aims to show how Pentaho Data Integration tool can be used to orchestrate machine learning workflow for London Air Quality prediction. The workflow includes data preparation, feature engineering, training models, updating models and deploying models to make predictions on new data streams. The diagram of the workflow is shown below. Description of the machine learning model used in the demo is provided in the next section.

The demo describes predictions of nitrogen dioxide concentration (NO2) at Harrow – Pinner Road site. NO2 is one of the key pollutant associated with the premature death in UK. The gas is largely created by diesel cars, lorries and buses, and affects lung capacity and growth. Predictions for other gases/sites can be performed in the equivalent manner.

## Machine Learning Orchestration Blueprint

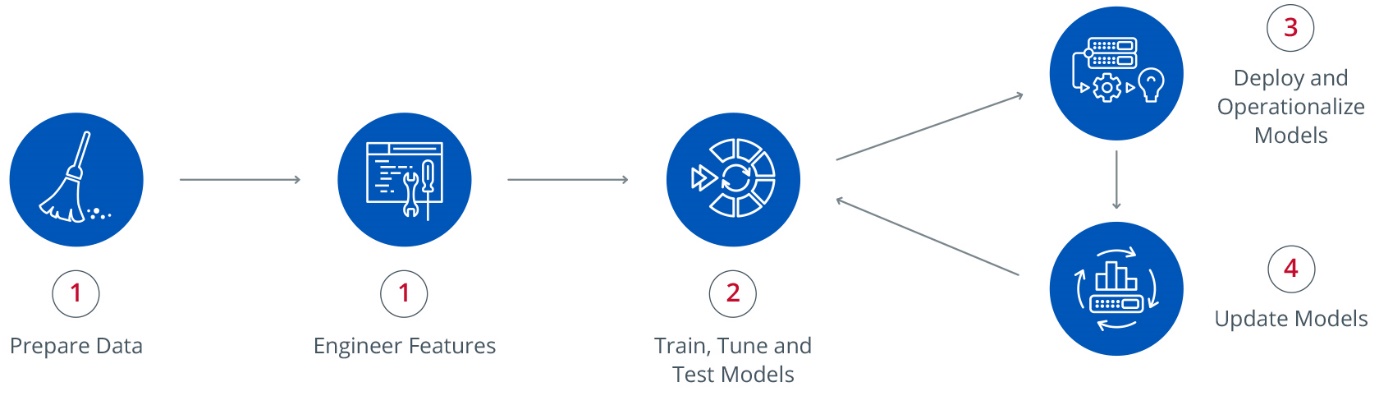


Figure 2. Machine Learning Workflow diagram.

## Dataset

The demo uses publicly available data published by Kings College London on air quality in London (https://www.londonair.org.uk/LondonAir/API/). The data provided by King’s College London includes hourly readings from sensor sites across London. Further details can be found in the LAQ workshop instructions document. The data is stored in Postgres database on AWS instance. Sensor readings are updated daily.

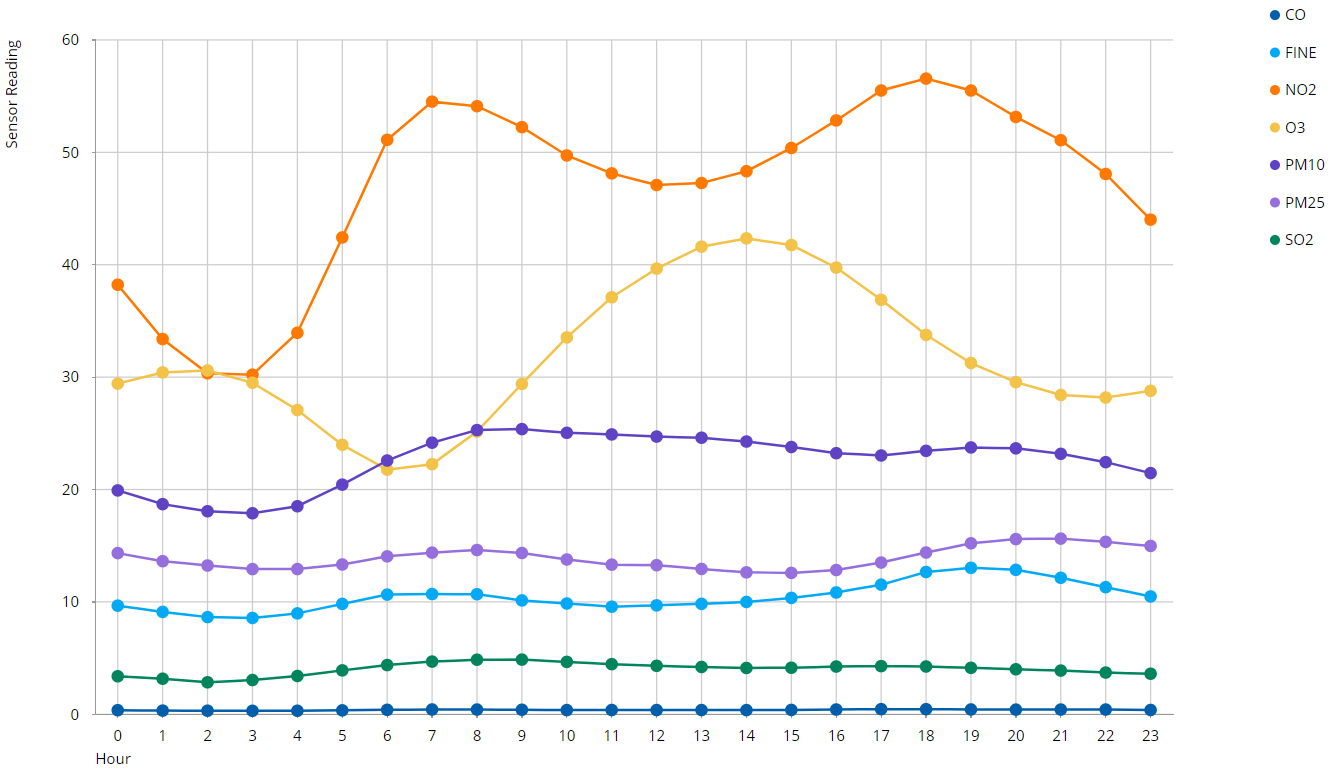


Figure 3. London Air Quality Data Example

## Model Used

To generate Air Quality Forecast, we used Long short-term memory (hereafter LSTM) model. LSTM is a type of Recurrent Neural Networks. LSTM architecture allows to learn patterns in the time series corresponding to short or long time periods. In this way, the network can capture trends, seasonal and daily patterns as well as episodic events. LSTM is a powerful technique for time series prediction and can be applied in various IoT/Predictive Use Cases. Links to additional materials on LSTM network can be found in section VI.

The model is configured to predict 12 hours ahead using 4 lagged features. Feature engineering and data preparation is described in more detail in section III.A. The model configuration is provided below (neurons = 4):

*model = Sequential()*

*model.add(LSTM(neurons, batch\_input\_shape=(batch\_size, X.shape[1], X.shape[2]), stateful=True))*

*model.add(Dense(1))*

*model.compile(loss='mean\_squared\_error', optimizer='adam')*

A picture containing object

Description generated with high confidence

Figure 4. LSTM model performance on test set.

## Prerequisites

In this demo “CPython Script Executor” step was used in Pentaho Data Integration workflow for model training and application on the new data (<http://wiki.pentaho.com/display/EAI/CPython+Script+Executor>).

The python version used is 3.5.2. The model was trained using Keras package (version 1.1.0) with Tensorflow backend

# Demo components

## Data Preparation

The data preparation and feature engineering for LSTM model training is performed in Pentaho Data Integration tool. The screen shot of the process is provided below. The process consists of reading the time series data, sorting it on date, generating columns with lagged data and removing nans.

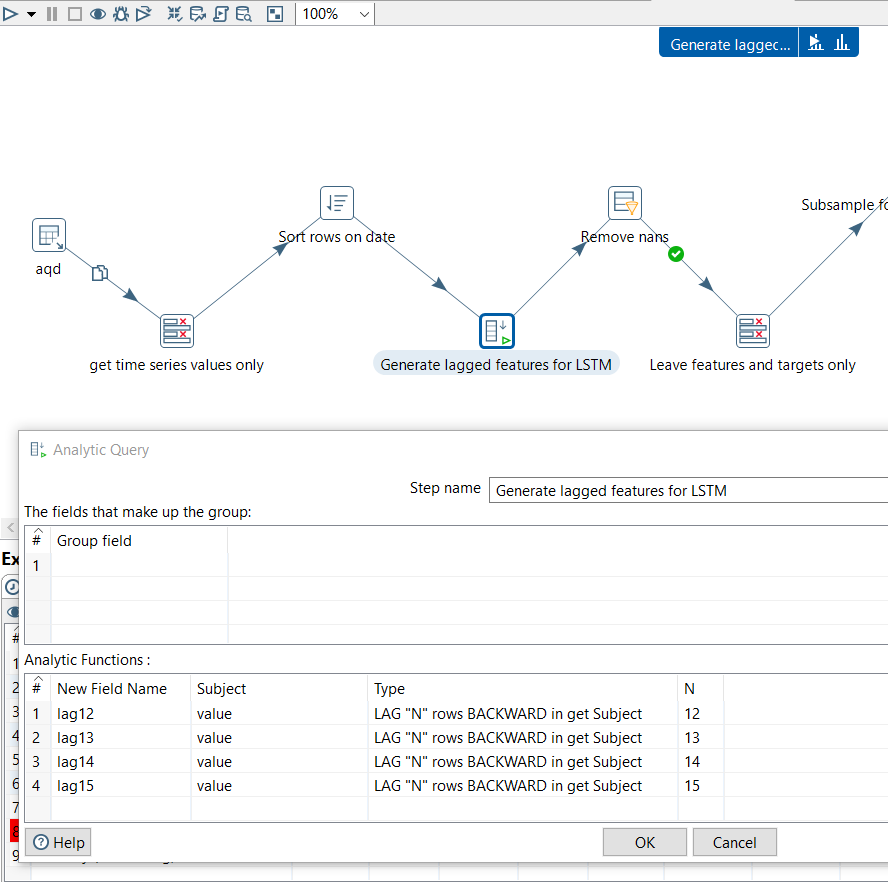


Figure 5. Data preparation workflow for model training in Pentaho Data Integration tool.

The data goes into the python script in the following form

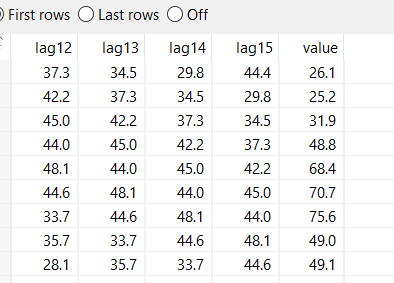
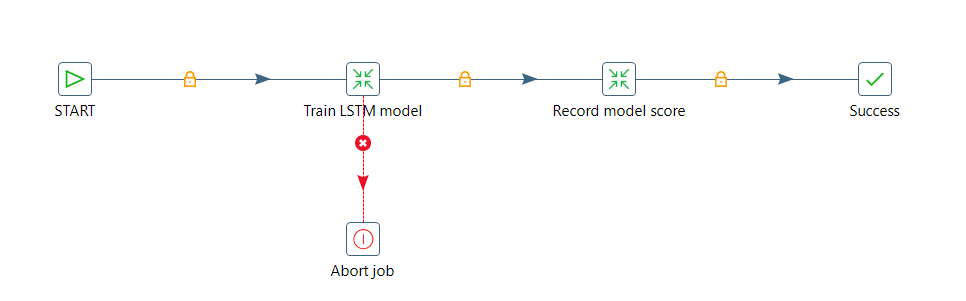


Figure 6. Format of data used for model training

, where “value” is a target variable in model training and “lag12”, “lag13”, “lag14”, “lag15” are sensor readings 12,13,14,15 hours before sensor reading “value” was observed.

## Model Training

After data is prepared in the right format, the model training is performed using python script embedded into ETL workflow. The model configuration and weights are saved during the script execution as well.



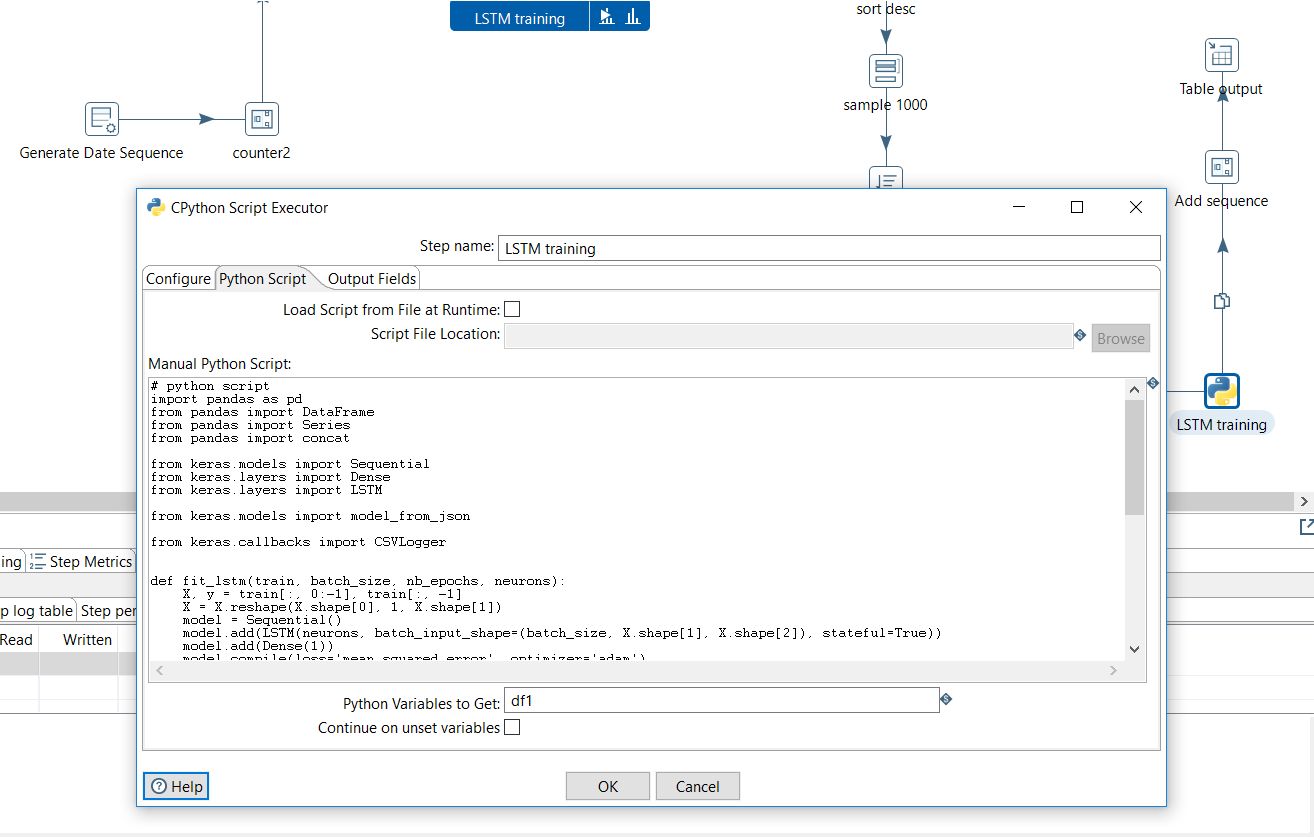
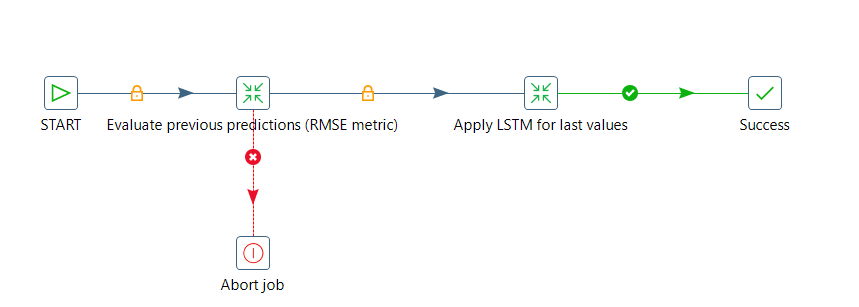


Figure 7. Python script in ETL workflow for model execution.

The complete workflow for model training can be found in “*LAQ\_LSTM\_train.kjb*”.

## Model Application

Based on the latest data obtained from the API the trained LSTM model generates predictions for the next 12 hours. The trained model is loaded via python script execution. The data is prepared for making predictions in ETL workflow. The generated predictions are stored in the database. The workflow for model application on new data can be found in “*LAQ\_LSTM\_apply\_on\_new\_data.kjb*”. The model application job is automated to run as new data becomes available from the source.



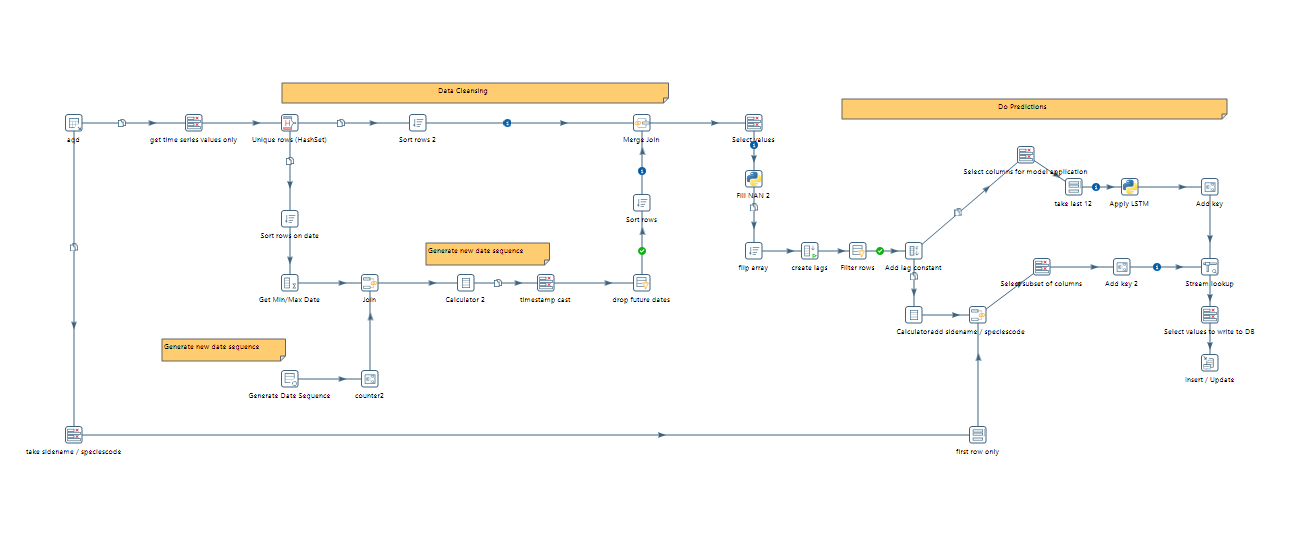


Figure 8. Pentaho workflow for generating predictions (12 hours ahead) based on the most recent sensor readings.

## Model Update

To update the model weights, the job for model training runs on the weekly basis. It ensures that the model is updated regularly as more data become available. As mentioned above new sensor readings are streamed from the API and stored in the database every night. Scheduling the job for model training entirely automates model update workflow. The full workflow can be found in *LAQ\_LSTM\_update\_the\_model.kjb*.

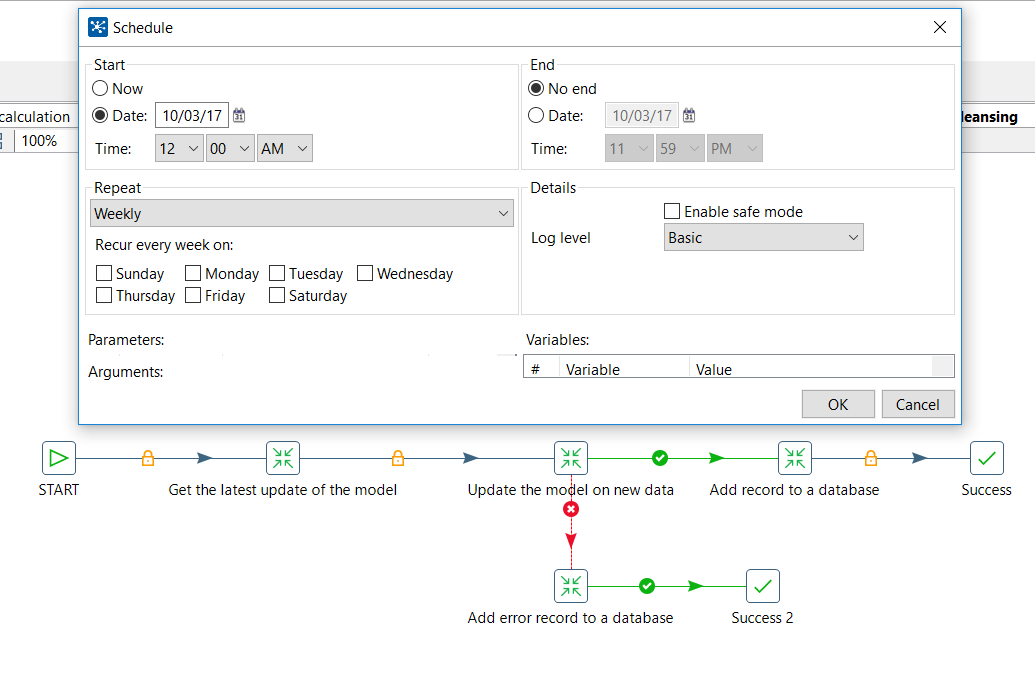


Figure 9. Scheduling model training job for regular update of model weights.

## Model Score

New data streamed from the API can be compared with predictions for the same period generated by the LSTM model. This comparison can be used to monitor the model performance over time and see how it improves with more data becoming available. As a score metric, RMSE was chosen. RMSE stands for Root Mean Square Error. The workflow for RMSE calculation can be found in “*RMSE\_calculation.ktr*”. The screenshot is also provided below

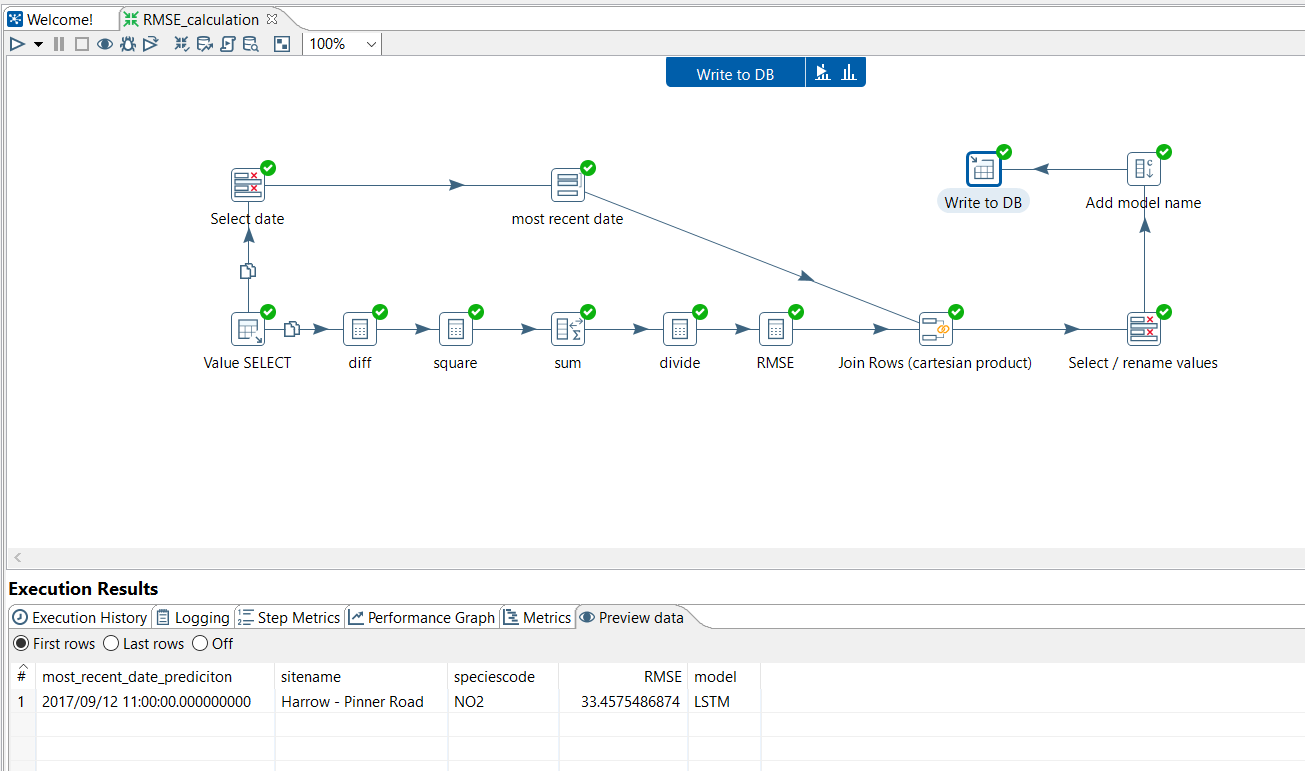


Figure 10. Pentaho workflow for model score evaluation.

## Dashboards

The generated by LSTM predictions and model scores are displayed in Pentaho dashboard. The dashboard also displays the loss curve for LSTM model training. As one can see loss is decreasing with epoch number indicating correct model architecture and training (*LAQ\_predictive\_dashboard.xdash*).

A close up of a map

Description generated with very high confidence

Figure 11. Pentaho Dashboard for prediction monitoring

# Demo Summary

The demo highlighted Pentaho end-to-end capabilities for machine learning workflow orchestration: preparing data, performing feature engineering, training a model, apply the trained model on new streamed data, updating the model and displaying predictions in Pentaho Dashboards. Pentaho help to automate the machine learning workflow and operationalize it for use in business operations.

A picture containing building

Description generated with high confidence

# Further Development

Including more features should improve the forecasts in future. A promising candidate is blending the sensor data with traffic data from the London Department for Transport. This data provides information about traffic volume in London and hence, allows for connecting air pollution with the actual count of cars and buses in the area. Unfortunately, these data are published once a year only and thus, cannot be used for real time forecasting.

Extending the demo for predictions in different areas and for different species is straight forward. This could be done either by implementing the parameters ‘authorityname’, ‘sidename’ and ‘speciescode’ as kettle variables or using meta data injection. That enables for forecasting different emission values with the same ktr. However, due to the specificity of a trained LSTM model, this required a pre-trained model for each combination of variables.

Different prediction models can be included in future as well, complementing the good performing, but training intensive LSTM network currently used. Through implementing the model as a CPython step, plugging in a different model is fairly easy. Further promising techniques include statistical models known from time series analysis (e.g. seasonal ARMA) as well as other recurrent network structures such as hierarchical temporal memory networks (HTM). Whereas the former is appropriate due to clear autocorrelation and a strong seasonal effect, is latter a promising new approach in forecasting of and anomaly detection in time series.

# Additional Reading

Recurrent Neural Networks/LSTM

<http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

<http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

General information about air pollution and related consequences

<http://www.londonair.org.uk/LondonAir/Default.aspx>

<http://www.telegraph.co.uk/science/2017/01/24/air-pollution-london-passes-levels-beijingand-wood-burners-making/>

<http://www.conserve-energy-future.com/causes-effects-solutions-of-air-pollution.php>

Pentaho Machine Learning Orchestration

<http://www.pentaho.com/machine-learning-orchestration>

Upcoming techniques for timeseries prediction:

<https://numenta.org/hierarchical-temporal-memory/>

<https://robjhyndman.com/talks/RevolutionR/10-Seasonal-ARIMA.pdf>

# Contact Details

Dr. Anya Rumyantseva

Email: [anna.rumyantseva@hitachivantara.com](mailto:anna.rumyantseva@hitachivantara.com)

Direct: +44-787-627-2032

Jonathan Doering

Email: [jonathan.doering@hitachivantara.com](mailto:jonathan.doering@hitachivantara.com)

Direct: +44-740-242-1522