MIE1628: Cloud-based Data Analytics
Assignment 5 – Tuesday, December 6th, 2022

Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca

Part A

Question 1: Explain below the 5 components shown in orange boxes. Explain which Azure components you will use where in this big data architecture and why.

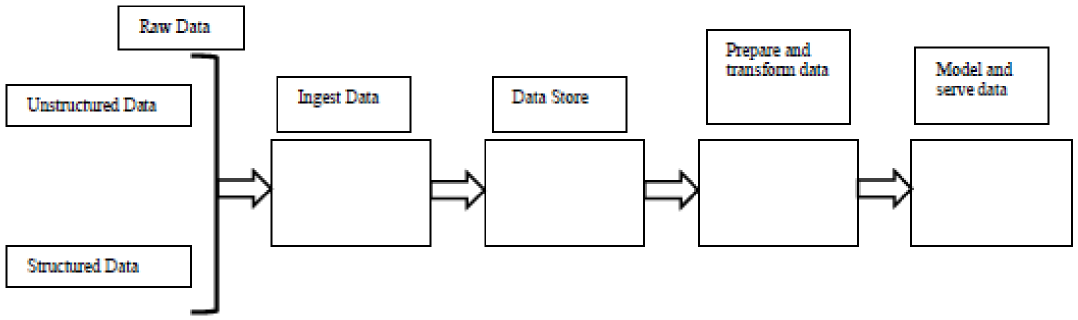
Azure Data Lake: Provides scalable storage for relational, semi-relational, and/or unstructured data

Azure Databricks: Data analytics platform optimized for Microsoft cloud services platform

Azure Data Factory: Cloud-based data integration service for Extract, Load, and Transform pipelines

Azure Synapse Analytics: Analytics engine designed to process big data quickly using parallel processing architecture

Azure Cosmos DB: Globally distributed, massively scalable, multi-model NoSQL database service for modern app development.



To ingest data: **Azure Data Factory** offers a scale-out, managed data movement solution. It allows users to ingest data with Azure Machine Learning.

Data store: **Azure Data Lake** allows the storage of any type of data of any size while Azure Cosmos DB only stores NoSQL data. Since the architecture deals with both structured and unstructured data, Azure Data Lake is used to store any ingested data.

Prepare and Transform Data: **Azure Databricks** allows the preparation and the transformation (i.e. clean, sort, merge, join, etc.) of data in a notebook environment. It is used to prepare and transform the stored data.

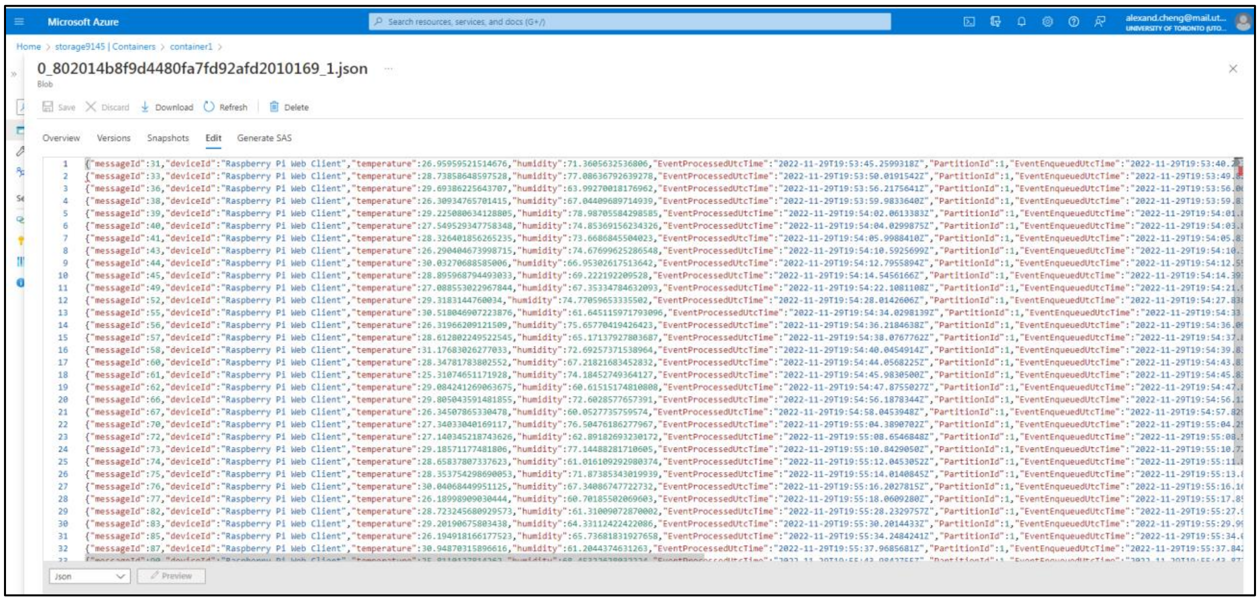
Model and serve data: **Azure Synapse Analytics** are a limitless analytics service able to fulfill the needs of complex big data analysis and modeling.

Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca

Question 2: Explain how Stream Analytics works in Azure.

Azure stream analytics can be used for real-time data ingestion and processing in the Azure cloud. Azure Stream Analytics allows the user to input (ingest) real time data, query (analyze), and an output (deliver). Data is ingested from Azure Event Hubs, IoT Hub, or Blob Storage. It allows simple transformation and queries to be applied on the data as well as machine learning. Analysis can be performed based on SQL query language for filtering, sorting, aggregating, and joining streaming data over a period of time. Transformed data can be sent to other services such as Azure Function, Service Bus, or event hubs, or can be sent to Power BI for visualization purposes and real-time dashboarding. Users can also store the data in other Azure storage services for training machine learning models.

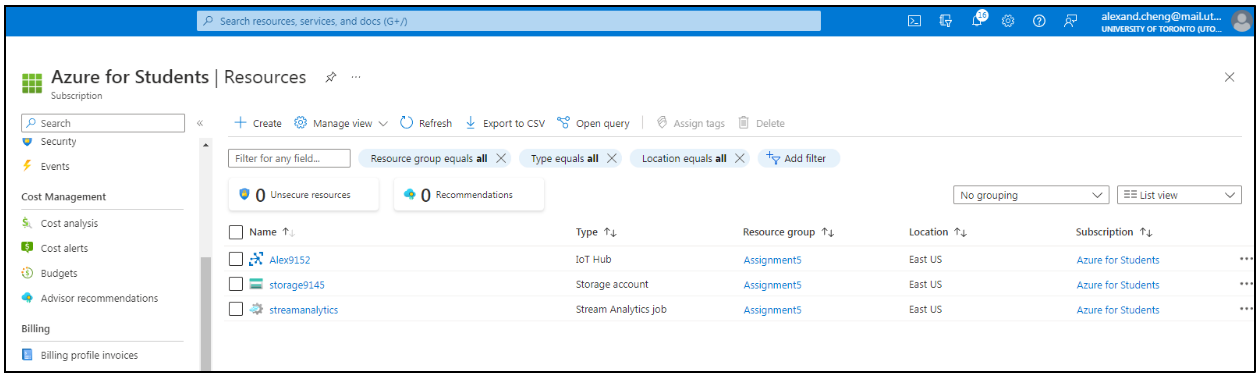
Question 3: Implementing a Stream Analytics job by using the Azure portal



0_802014b8f9d4480fa7fd92afd2010169_1.json
Blob

```
{"messageId":31,"deviceId":"Raspberry Pi Web Client","temperature":26.95959521514676,"humidity":71.3605632536806,"EventProcessedUtcTime":"2022-11-29T19:53:45.2599318Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:53:40.
{"messageId":33,"deviceId":"Raspberry Pi Web Client","temperature":28.73858648597528,"humidity":77.08636792639278,"EventProcessedUtcTime":"2022-11-29T19:53:50.0191542Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:53:49.
{"messageId":36,"deviceId":"Raspberry Pi Web Client","temperature":29.69386225643707,"humidity":63.99270018176962,"EventProcessedUtcTime":"2022-11-29T19:53:56.2175641Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:53:56.
{"messageId":38,"deviceId":"Raspberry Pi Web Client","temperature":26.30934765701415,"humidity":67.04409689714939,"EventProcessedUtcTime":"2022-11-29T19:53:59.9833640Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:53:59.
{"messageId":39,"deviceId":"Raspberry Pi Web Client","temperature":29.225080634128805,"humidity":78.98705584298585,"EventProcessedUtcTime":"2022-11-29T19:54:02.0613383Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:01.
{"messageId":40,"deviceId":"Raspberry Pi Web Client","temperature":27.549529347758348,"humidity":74.85369156234326,"EventProcessedUtcTime":"2022-11-29T19:54:04.0298875Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:03.
{"messageId":41,"deviceId":"Raspberry Pi Web Client","temperature":28.326401856265235,"humidity":73.6686845504023,"EventProcessedUtcTime":"2022-11-29T19:54:05.9988410Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:05.
{"messageId":43,"deviceId":"Raspberry Pi Web Client","temperature":26.290404673998715,"humidity":74.67699625286548,"EventProcessedUtcTime":"2022-11-29T19:54:10.5925699Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:10.
{"messageId":44,"deviceId":"Raspberry Pi Web Client","temperature":30.03270688585006,"humidity":66.95302617513642,"EventProcessedUtcTime":"2022-11-29T19:54:12.7955894Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:12.
{"messageId":45,"deviceId":"Raspberry Pi Web Client","temperature":28.895968794493033,"humidity":69.222192209528,"EventProcessedUtcTime":"2022-11-29T19:54:14.5456166Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:14.
{"messageId":49,"deviceId":"Raspberry Pi Web Client","temperature":27.088553022967844,"humidity":67.35334784632093,"EventProcessedUtcTime":"2022-11-29T19:54:22.1081108Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:21.
{"messageId":52,"deviceId":"Raspberry Pi Web Client","temperature":29.3183144760034,"humidity":74.77059653335502,"EventProcessedUtcTime":"2022-11-29T19:54:28.0142606Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:27.
{"messageId":55,"deviceId":"Raspberry Pi Web Client","temperature":30.518046907223876,"humidity":61.645115971793096,"EventProcessedUtcTime":"2022-11-29T19:54:34.0298139Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:33.
{"messageId":56,"deviceId":"Raspberry Pi Web Client","temperature":26.31966209121509,"humidity":75.65770419426423,"EventProcessedUtcTime":"2022-11-29T19:54:36.2184638Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:36.
{"messageId":57,"deviceId":"Raspberry Pi Web Client","temperature":28.612802249522545,"humidity":65.17137927803687,"EventProcessedUtcTime":"2022-11-29T19:54:38.0767762Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:37.
{"messageId":58,"deviceId":"Raspberry Pi Web Client","temperature":31.17683026277033,"humidity":72.69257371538964,"EventProcessedUtcTime":"2022-11-29T19:54:40.0454914Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:39.
{"messageId":60,"deviceId":"Raspberry Pi Web Client","temperature":28.34781783802552,"humidity":67.21821683452832,"EventProcessedUtcTime":"2022-11-29T19:54:44.0568225Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:43.
{"messageId":61,"deviceId":"Raspberry Pi Web Client","temperature":25.31074651171928,"humidity":74.18452749364127,"EventProcessedUtcTime":"2022-11-29T19:54:45.9830500Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:45.
{"messageId":62,"deviceId":"Raspberry Pi Web Client","temperature":29.084241269063675,"humidity":60.61515174810808,"EventProcessedUtcTime":"2022-11-29T19:54:47.8755027Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:47.
{"messageId":66,"deviceId":"Raspberry Pi Web Client","temperature":29.805043591481855,"humidity":72.6028577657391,"EventProcessedUtcTime":"2022-11-29T19:54:56.1878344Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:56.
{"messageId":67,"deviceId":"Raspberry Pi Web Client","temperature":26.34507865330478,"humidity":60.0527735759574,"EventProcessedUtcTime":"2022-11-29T19:54:58.0453948Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:54:57.
{"messageId":70,"deviceId":"Raspberry Pi Web Client","temperature":27.34033040169117,"humidity":76.50476186277967,"EventProcessedUtcTime":"2022-11-29T19:55:04.3890702Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:55:04.
{"messageId":72,"deviceId":"Raspberry Pi Web Client","temperature":27.140345218743626,"humidity":62.89182693230172,"EventProcessedUtcTime":"2022-11-29T19:55:08.6546848Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:55:08.
{"messageId":73,"deviceId":"Raspberry Pi Web Client","temperature":29.18571177481806,"humidity":77.14488281710605,"EventProcessedUtcTime":"2022-11-29T19:55:10.8429050Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:55:10.
{"messageId":74,"deviceId":"Raspberry Pi Web Client","temperature":28.65837807337623,"humidity":61.016109292980374,"EventProcessedUtcTime":"2022-11-29T19:55:12.0453052Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:55:11.
{"messageId":75,"deviceId":"Raspberry Pi Web Client","temperature":28.353754298690053,"humidity":71.87385343019939,"EventProcessedUtcTime":"2022-11-29T19:55:14.0140845Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:55:13.
{"messageId":76,"deviceId":"Raspberry Pi Web Client","temperature":30.04068449951125,"humidity":67.34086747722732,"EventProcessedUtcTime":"2022-11-29T19:55:16.2027815Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:55:16.
{"messageId":77,"deviceId":"Raspberry Pi Web Client","temperature":26.18998909030444,"humidity":60.70185502069603,"EventProcessedUtcTime":"2022-11-29T19:55:18.0609280Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:55:17.
{"messageId":82,"deviceId":"Raspberry Pi Web Client","temperature":28.723245680929573,"humidity":61.31009072870002,"EventProcessedUtcTime":"2022-11-29T19:55:28.2329757Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:55:27.
{"messageId":83,"deviceId":"Raspberry Pi Web Client","temperature":29.20190675803438,"humidity":64.33112422422086,"EventProcessedUtcTime":"2022-11-29T19:55:30.2014433Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:55:29.
{"messageId":85,"deviceId":"Raspberry Pi Web Client","temperature":26.194918166177523,"humidity":65.73681831927658,"EventProcessedUtcTime":"2022-11-29T19:55:34.2484241Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:55:34.
{"messageId":87,"deviceId":"Raspberry Pi Web Client","temperature":30.94870315896616,"humidity":61.2044374631263,"EventProcessedUtcTime":"2022-11-29T19:55:37.9685681Z","PartitionId":1,"EventEnqueuedUtcTime":"2022-11-29T19:55:37.
```

Screenshot of the top 30 results

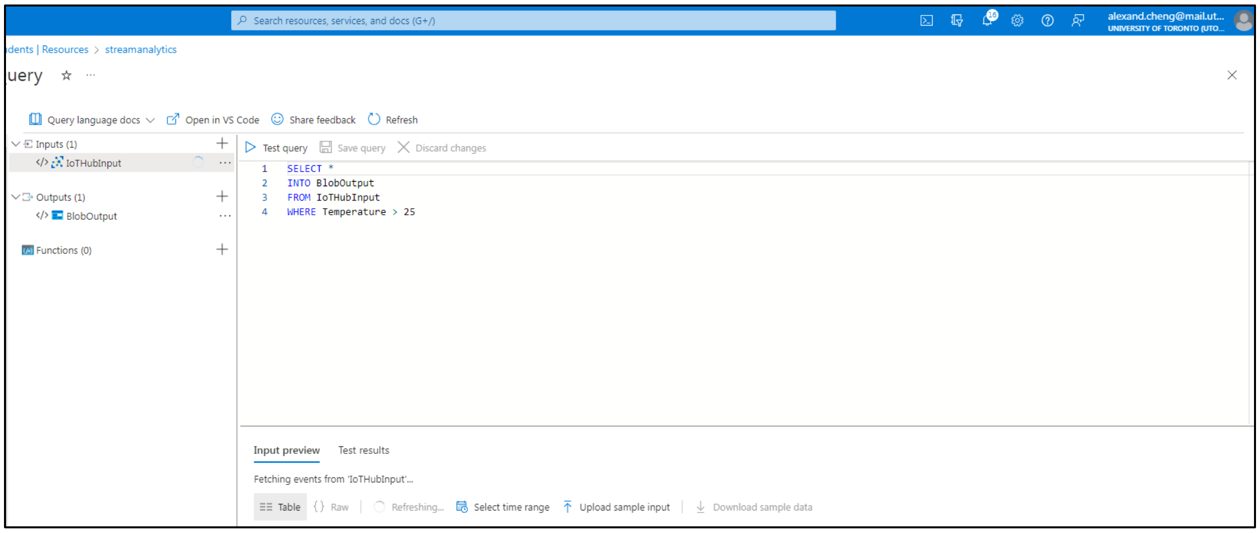


Azure for Students | Resources

Name	Type	Resource group	Location	Subscription
Alex9152	IoT Hub	Assignment5	East US	Azure for Students
storage9145	Storage account	Assignment5	East US	Azure for Students
streamanalytics	Stream Analytics job	Assignment5	East US	Azure for Students

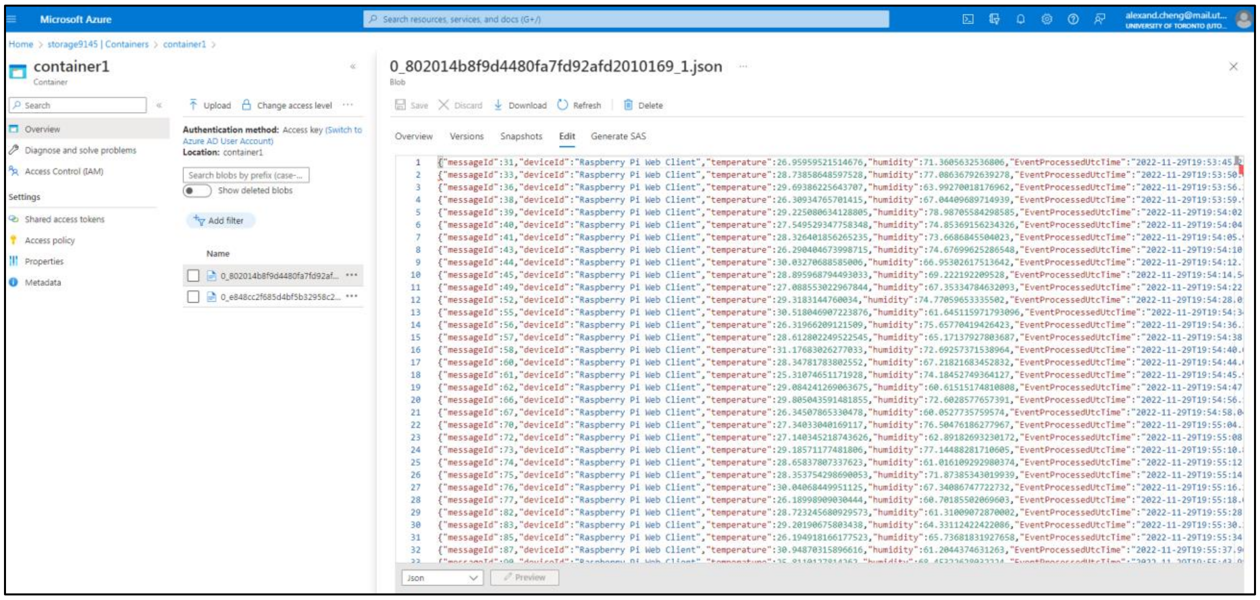
Deployed Resources

Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca



```
SELECT *
INTO BlobOutput
FROM IoTHubInput
WHERE Temperature > 25
```

Screenshot of Query



```
{"messageId":31,"deviceId":"Raspberry Pi Web Client","temperature":26.95959521514676,"humidity":71.3605632536806,"EventProcessedUtcTime":"2022-11-29T19:53:45
{"messageId":33,"deviceId":"Raspberry Pi Web Client","temperature":28.73858648597528,"humidity":77.08636792639278,"EventProcessedUtcTime":"2022-11-29T19:53:50
{"messageId":36,"deviceId":"Raspberry Pi Web Client","temperature":29.69386225643707,"humidity":63.99270018176962,"EventProcessedUtcTime":"2022-11-29T19:53:56.
{"messageId":38,"deviceId":"Raspberry Pi Web Client","temperature":26.30934765701415,"humidity":67.04409689714939,"EventProcessedUtcTime":"2022-11-29T19:53:59.
{"messageId":39,"deviceId":"Raspberry Pi Web Client","temperature":29.225080634128805,"humidity":78.98705584298585,"EventProcessedUtcTime":"2022-11-29T19:54:02
{"messageId":40,"deviceId":"Raspberry Pi Web Client","temperature":27.549529347758348,"humidity":74.85369156234326,"EventProcessedUtcTime":"2022-11-29T19:54:04
{"messageId":41,"deviceId":"Raspberry Pi Web Client","temperature":28.326401856265235,"humidity":73.6686845504023,"EventProcessedUtcTime":"2022-11-29T19:54:05.
{"messageId":43,"deviceId":"Raspberry Pi Web Client","temperature":26.290404673998715,"humidity":74.67699625286548,"EventProcessedUtcTime":"2022-11-29T19:54:10
{"messageId":44,"deviceId":"Raspberry Pi Web Client","temperature":30.03270688585006,"humidity":66.95302617513642,"EventProcessedUtcTime":"2022-11-29T19:54:12.
{"messageId":45,"deviceId":"Raspberry Pi Web Client","temperature":28.895968794493033,"humidity":69.222192209528,"EventProcessedUtcTime":"2022-11-29T19:54:14.5
{"messageId":49,"deviceId":"Raspberry Pi Web Client","temperature":27.088553022967844,"humidity":67.35334784632093,"EventProcessedUtcTime":"2022-11-29T19:54:22
{"messageId":52,"deviceId":"Raspberry Pi Web Client","temperature":29.3183144760034,"humidity":74.77059653335502,"EventProcessedUtcTime":"2022-11-29T19:54:28.0
{"messageId":55,"deviceId":"Raspberry Pi Web Client","temperature":30.518046907223876,"humidity":61.645115971793096,"EventProcessedUtcTime":"2022-11-29T19:54:3
{"messageId":56,"deviceId":"Raspberry Pi Web Client","temperature":26.31966209121509,"humidity":75.65770419426423,"EventProcessedUtcTime":"2022-11-29T19:54:36.
{"messageId":57,"deviceId":"Raspberry Pi Web Client","temperature":28.612802249522545,"humidity":65.17137927803687,"EventProcessedUtcTime":"2022-11-29T19:54:38
{"messageId":58,"deviceId":"Raspberry Pi Web Client","temperature":31.17683026277033,"humidity":72.69257371538964,"EventProcessedUtcTime":"2022-11-29T19:54:40.
{"messageId":60,"deviceId":"Raspberry Pi Web Client","temperature":28.34781783802552,"humidity":67.21821683452832,"EventProcessedUtcTime":"2022-11-29T19:54:44.
{"messageId":61,"deviceId":"Raspberry Pi Web Client","temperature":25.31074651171928,"humidity":74.18452749364127,"EventProcessedUtcTime":"2022-11-29T19:54:45.
{"messageId":62,"deviceId":"Raspberry Pi Web Client","temperature":29.084241269063675,"humidity":60.61515174810808,"EventProcessedUtcTime":"2022-11-29T19:54:47
{"messageId":66,"deviceId":"Raspberry Pi Web Client","temperature":29.805043591481855,"humidity":72.6028577657391,"EventProcessedUtcTime":"2022-11-29T19:54:56.
{"messageId":67,"deviceId":"Raspberry Pi Web Client","temperature":26.34507865330478,"humidity":60.0527735759574,"EventProcessedUtcTime":"2022-11-29T19:54:58.0
{"messageId":70,"deviceId":"Raspberry Pi Web Client","temperature":27.34033040169117,"humidity":76.50476186277967,"EventProcessedUtcTime":"2022-11-29T19:55:04.
{"messageId":72,"deviceId":"Raspberry Pi Web Client","temperature":27.140345218743626,"humidity":62.89182693230172,"EventProcessedUtcTime":"2022-11-29T19:55:08
{"messageId":73,"deviceId":"Raspberry Pi Web Client","temperature":29.18571177481806,"humidity":77.14488281710605,"EventProcessedUtcTime":"2022-11-29T19:55:10.
{"messageId":74,"deviceId":"Raspberry Pi Web Client","temperature":28.65837807337623,"humidity":61.016109292980374,"EventProcessedUtcTime":"2022-11-29T19:55:12
{"messageId":75,"deviceId":"Raspberry Pi Web Client","temperature":28.353754298690053,"humidity":71.87385343019939,"EventProcessedUtcTime":"2022-11-29T19:55:14
{"messageId":76,"deviceId":"Raspberry Pi Web Client","temperature":30.04068449951125,"humidity":67.34086747722732,"EventProcessedUtcTime":"2022-11-29T19:55:16.
{"messageId":77,"deviceId":"Raspberry Pi Web Client","temperature":26.18998909030444,"humidity":60.70185502069603,"EventProcessedUtcTime":"2022-11-29T19:55:18.
{"messageId":82,"deviceId":"Raspberry Pi Web Client","temperature":28.723245680929573,"humidity":61.31009072870002,"EventProcessedUtcTime":"2022-11-29T19:55:28
{"messageId":83,"deviceId":"Raspberry Pi Web Client","temperature":29.20190675803438,"humidity":64.33112422422086,"EventProcessedUtcTime":"2022-11-29T19:55:30.
{"messageId":85,"deviceId":"Raspberry Pi Web Client","temperature":26.194918166177523,"humidity":65.73681831927658,"EventProcessedUtcTime":"2022-11-29T19:55:34
{"messageId":87,"deviceId":"Raspberry Pi Web Client","temperature":30.94870315896616,"humidity":61.2044374631263,"EventProcessedUtcTime":"2022-11-29T19:55:37.9
```

Another screenshot of the top 30 results

Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca

Part B – Implement in Azure Machine Learning using Azure Notebook

https://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data+Set

1. Explain what problem you are going to solve using this dataset. Provide a brief overview of your problem statement. [Discuss your problem statement with your TAs if they approve then you can proceed with the next steps.]

Problem Statement:

The gas turbine is very useful in the modern world since it is relatively compact in size and creates a lot of power. Gas turbines are used in backup power systems in Manhattan, for example, when the grid goes down due to natural disasters, gas turbines power up and can produce power for emergencies.

The dataset contains 36733 instances of 11 sensor measures aggregated over one hour (by means of average or sum) from a gas turbine located in Turkey's northwestern region for the purpose of studying flue gas emissions, namely CO and NOx (NO + NO2). Some of the features listed are ambient temperature, pressure, humidity, etc. We can examine which main features are used for predicting hourly net energy yield.

By determining which features are the most significant in determining the Turbine energy yield, we can determine if the environment is suitable for the energy yield required (clients will have a minimum energy output needed if used as a backup), this can save a lot of cost by researching the environment and critical components before implementing and finding out that the location chosen is not suitable. By comparing experimental results with theoretical results for gas turbine energy yield, the user can determine if there's any external factors that are significant that are not being included.

There will be no data cleaning required since there are no N/A or missing values inside the dataset, however, the data is separated by years throughout 5 different years. We will merge the first three years to be used as our training data and the last two years will be used as our test data to determine how accurate our model is, this will be a 60/40 split. Since the data does not require data cleaning, more effort will be placed on exploratory data analysis and modelling. We will spend a portion on examining the gas emissions (CO and NOx), due to the rising concerns of climate change and renewable energy.

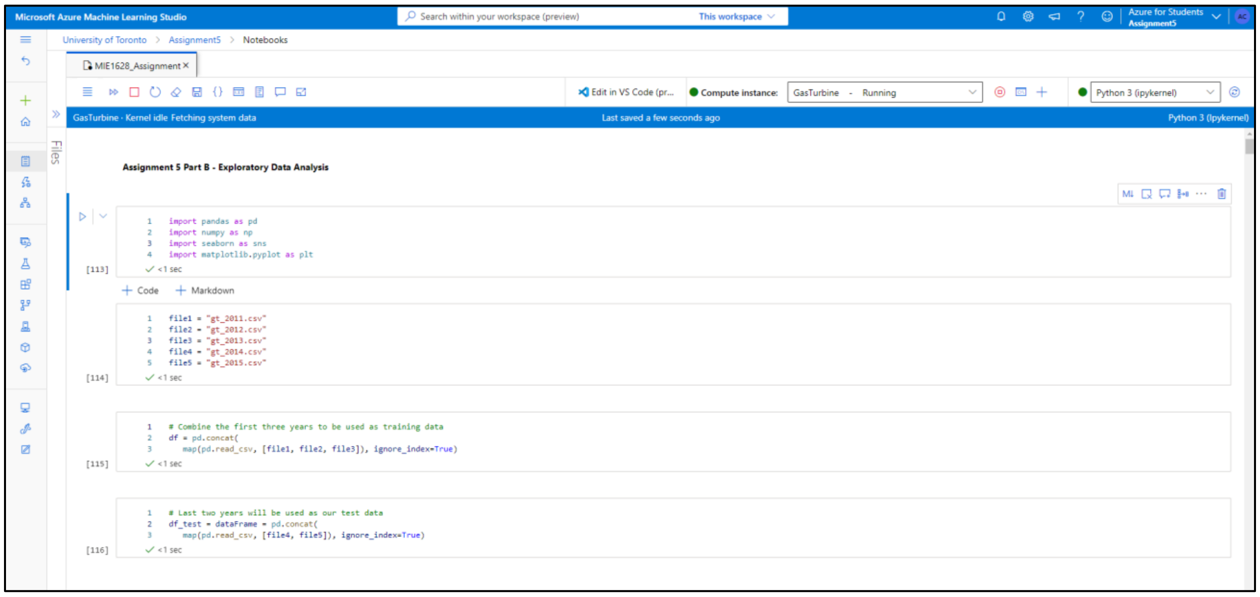
Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca

2. Explain your dataset. Explore your dataset and provide at least 5 meaningful charts/graphs with an explanation.
3. Do data cleaning/pre-processing as required and explain what you have done for your dataset and why?

A summary of the Exploratory Data Analysis and screenshots will be shown: Please look at the ipynb file for a detailed explanation of each step, otherwise there will be too many screenshots.

List of Exploratory Data Analysis for the Dataset

1. Correlation Plot - Can see which features are related with one another
2. Histogram - see the count distribution of each feature
3. Plot each feature against TEY to see the direct relationship
4. Pairplot to see relationship between all features
5. Time Series to compare changes/differences between years for TEY and CO (Climate Change) to see significance differences


Microsoft Azure Machine Learning Studio
Search within your workspace (preview)
This workspace
Azure for Students
Assignment5
University of Toronto > Assignment5 > Notebooks
MIE1628_Assignment
Edit in VS Code (pr...
Compute instance: GasTurbine - Running
Python 3 (ipykernel)
GasTurbine · Kernel idle Fetching system data
Last saved a few seconds ago
Python 3 (ipykernel)
Files
+ Code
+ Markdown

Assignment 5 Part B - Exploratory Data Analysis

```
[113]
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
✓ <1 sec
```

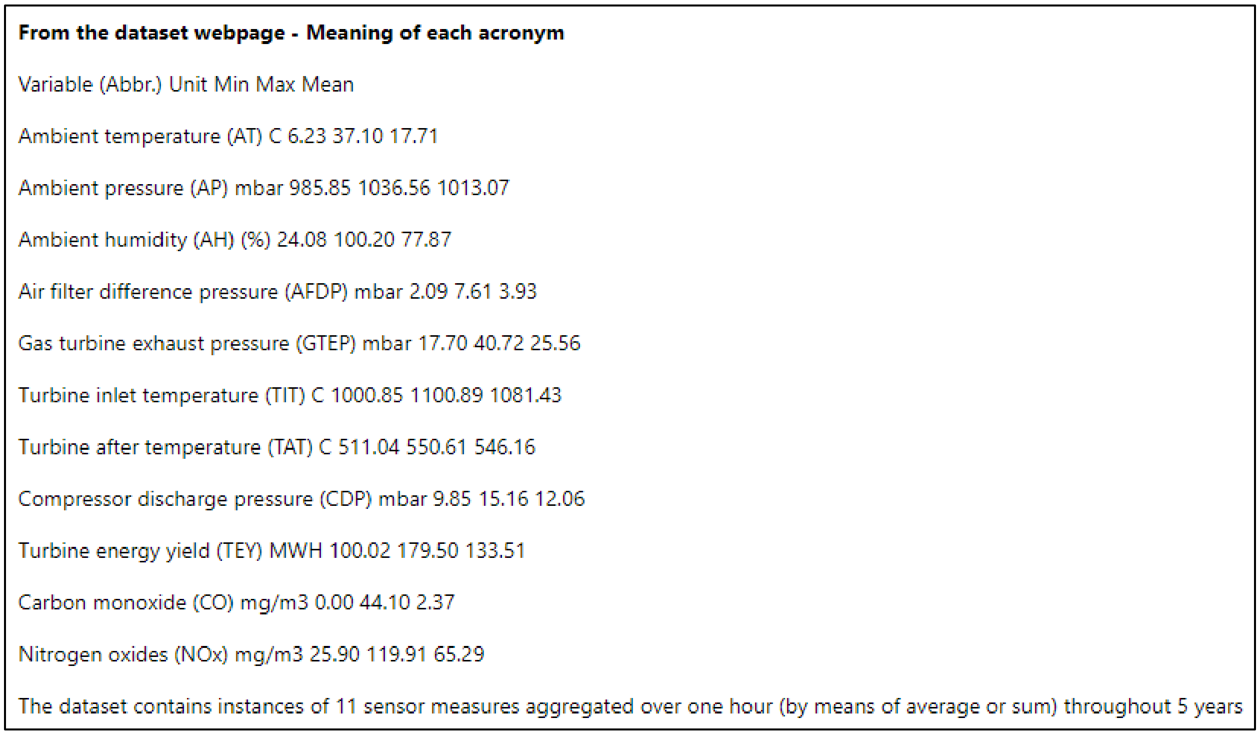
```
[114]
file1 = "gt_2011.csv"
file2 = "gt_2012.csv"
file3 = "gt_2013.csv"
file4 = "gt_2014.csv"
file5 = "gt_2015.csv"
✓ <1 sec
```

```
[115]
# Combine the first three years to be used as training data
df = pd.concat(
   map(pd.read_csv, [file1, file2, file3]), ignore_index=True)
✓ <1 sec
```

```
[116]
# Last two years will be used as our test data
df_test = dataFrame = pd.concat(
   map(pd.read_csv, [file4, file5]), ignore_index=True)
✓ <1 sec
```

Deployed compute instance to run notebook on Azure Portal

Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca



From the dataset webpage - Meaning of each acronym

Variable (Abbr.) Unit Min Max Mean

Ambient temperature (AT) C 6.23 37.10 17.71

Ambient pressure (AP) mbar 985.85 1036.56 1013.07

Ambient humidity (AH) (%) 24.08 100.20 77.87

Air filter difference pressure (AFDP) mbar 2.09 7.61 3.93

Gas turbine exhaust pressure (GTEP) mbar 17.70 40.72 25.56

Turbine inlet temperature (TIT) C 1000.85 1100.89 1081.43

Turbine after temperature (TAT) C 511.04 550.61 546.16

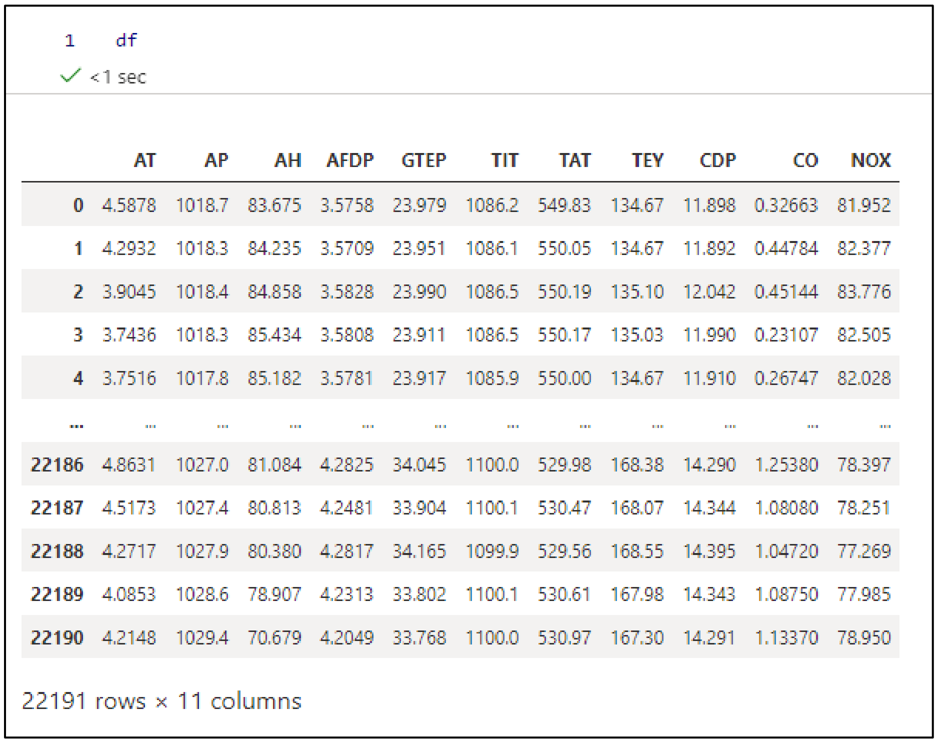
Compressor discharge pressure (CDP) mbar 9.85 15.16 12.06

Turbine energy yield (TEY) MWH 100.02 179.50 133.51

Carbon monoxide (CO) mg/m3 0.00 44.10 2.37

Nitrogen oxides (NOx) mg/m3 25.90 119.91 65.29

The dataset contains instances of 11 sensor measures aggregated over one hour (by means of average or sum) throughout 5 years



```
df
```

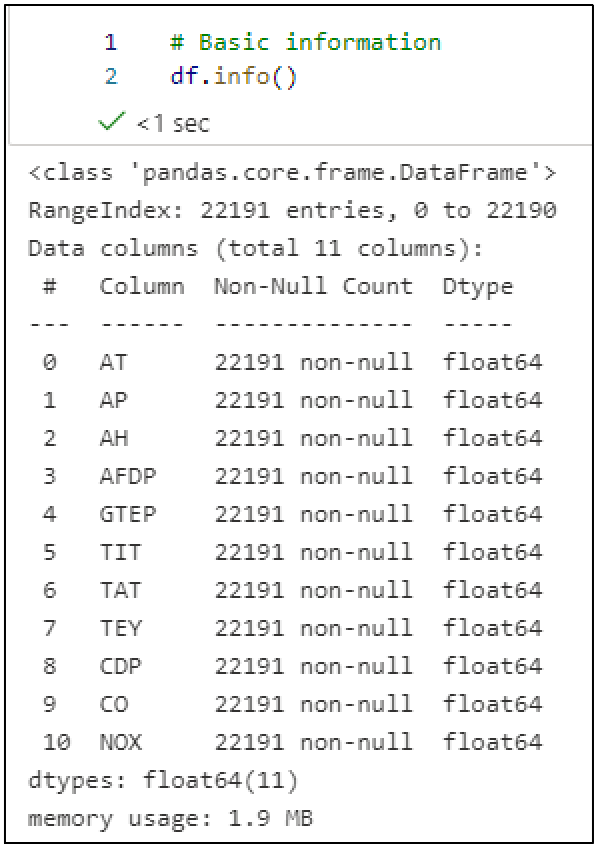
✓ <1 sec

	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
0	4.5878	1018.7	83.675	3.5758	23.979	1086.2	549.83	134.67	11.898	0.32663	81.952
1	4.2932	1018.3	84.235	3.5709	23.951	1086.1	550.05	134.67	11.892	0.44784	82.377
2	3.9045	1018.4	84.858	3.5828	23.990	1086.5	550.19	135.10	12.042	0.45144	83.776
3	3.7436	1018.3	85.434	3.5808	23.911	1086.5	550.17	135.03	11.990	0.23107	82.505
4	3.7516	1017.8	85.182	3.5781	23.917	1085.9	550.00	134.67	11.910	0.26747	82.028
...	...	...	...	...	...	...	...	...	...	...	...
22186	4.8631	1027.0	81.084	4.2825	34.045	1100.0	529.98	168.38	14.290	1.25380	78.397
22187	4.5173	1027.4	80.813	4.2481	33.904	1100.1	530.47	168.07	14.344	1.08080	78.251
22188	4.2717	1027.9	80.380	4.2817	34.165	1099.9	529.56	168.55	14.395	1.04720	77.269
22189	4.0853	1028.6	78.907	4.2313	33.802	1100.1	530.61	167.98	14.343	1.08750	77.985
22190	4.2148	1029.4	70.679	4.2049	33.768	1100.0	530.97	167.30	14.291	1.13370	78.950

22191 rows × 11 columns

Training Data, the target variable is the Turbine Energy Yield (TEY)

Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca



```
# Basic information
df.info()
```

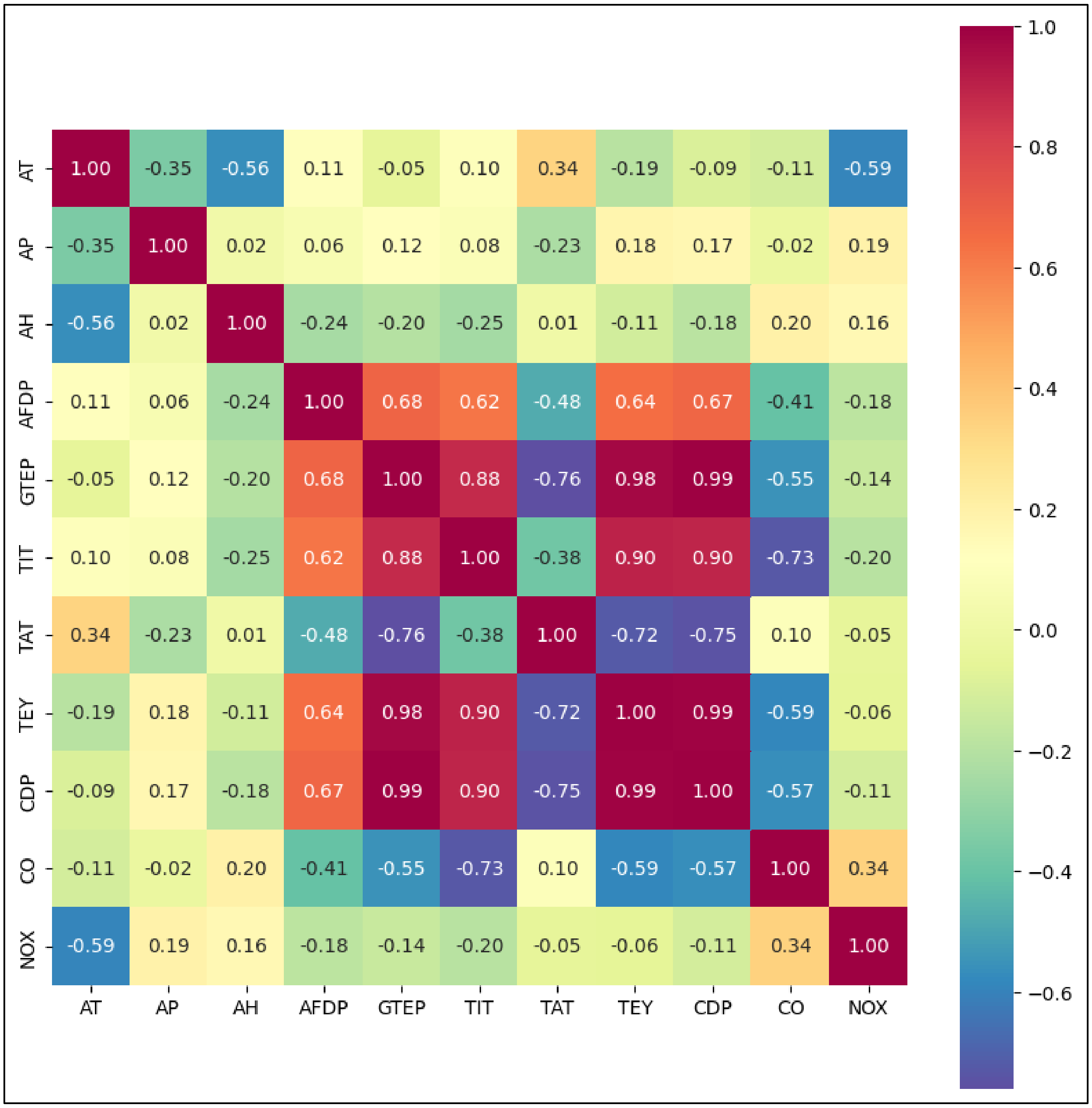
✓ <1 sec

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22191 entries, 0 to 22190
Data columns (total 11 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   AT      22191 non-null  float64
 1   AP      22191 non-null  float64
 2   AH      22191 non-null  float64
 3   AFDP    22191 non-null  float64
 4   GTEP    22191 non-null  float64
 5   TIT     22191 non-null  float64
 6   TAT     22191 non-null  float64
 7   TEY     22191 non-null  float64
 8   CDP     22191 non-null  float64
 9   CO      22191 non-null  float64
 10  NOX     22191 non-null  float64
dtypes: float64(11)
memory usage: 1.9 MB
```

We can observe that our data does not require any data cleaning/pre-processing

Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca

Correlation Plot

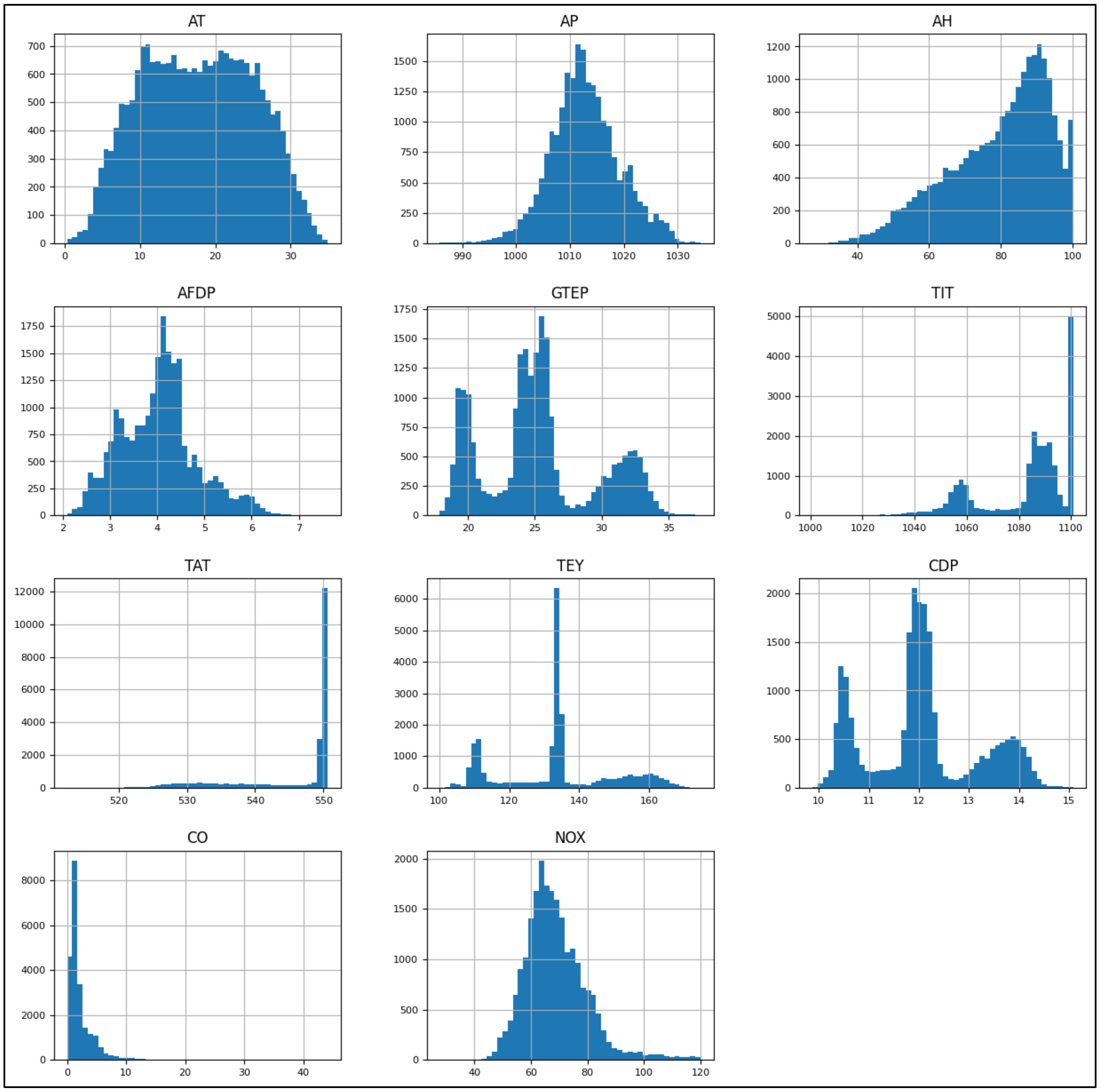


We can observe that multiple features are correlated with one another, mostly the compressor features and not regarding the environment conditions.

Features such as the air filter difference pressure, gas turbine exhaust pressure, turbine inlet temperature, turbine energy yield, and compressor discharge pressure are highly correlated with one another.

For the Turbine energy yield, we can observe that the values are highly correlated with the gas turbine exhaust pressure, turbine inlet temperature, and compressor discharge pressure.

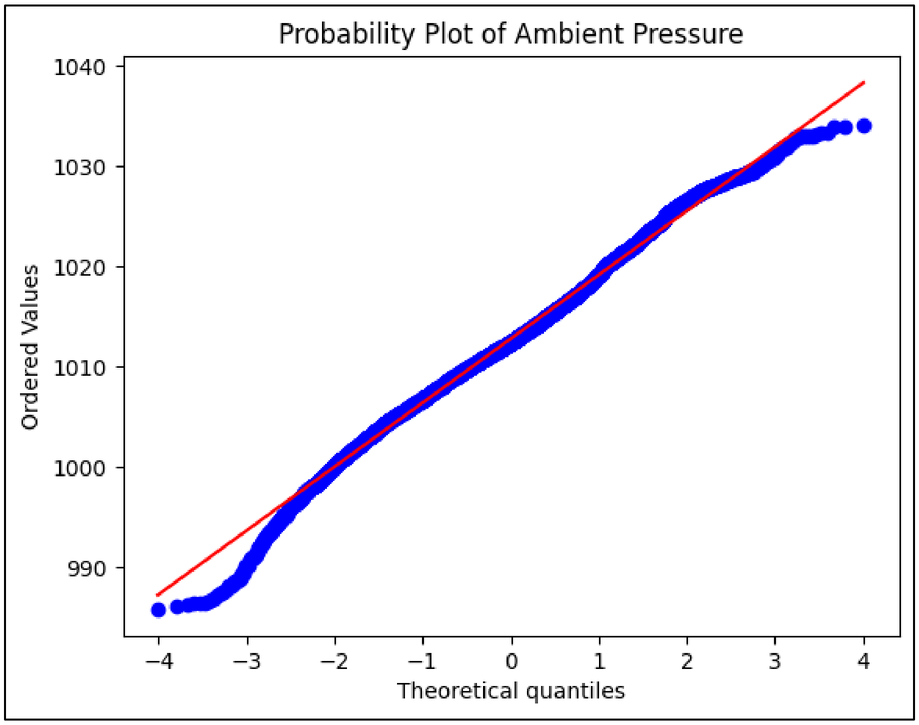
Histogram

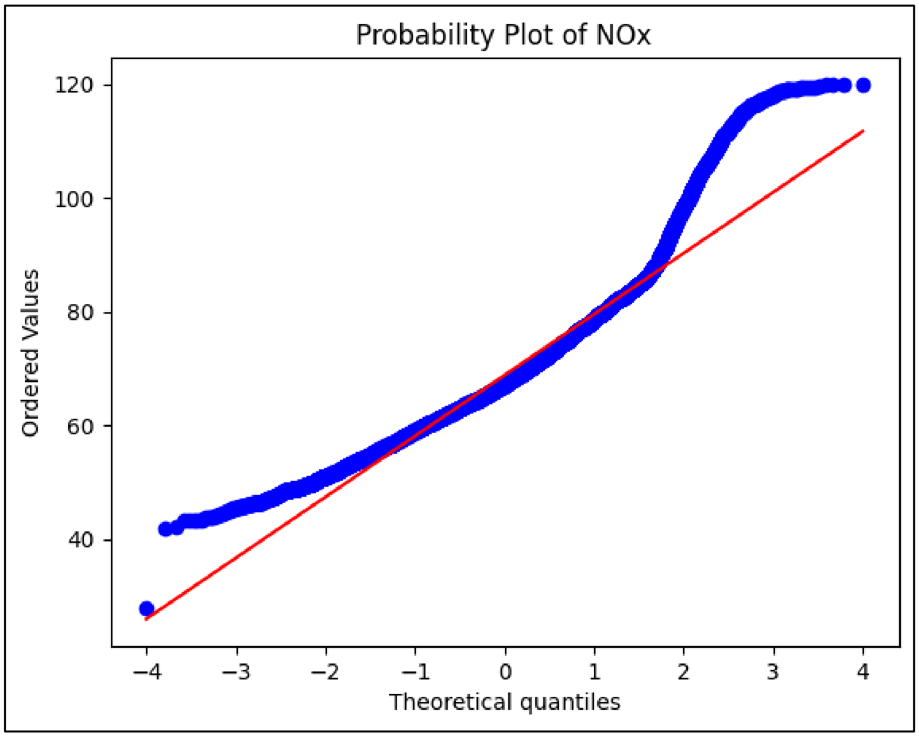


We can observe that there are some features that have a similar distribution to TEY such as the Compressor Discharge Pressure (CDP) and a bit of Gas Turbine Exhaust Pressure (GTEP), we can also observe that Ambient Pressure (AP) and NOx follow like a normal distribution based on the bell-shaped curve

Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca

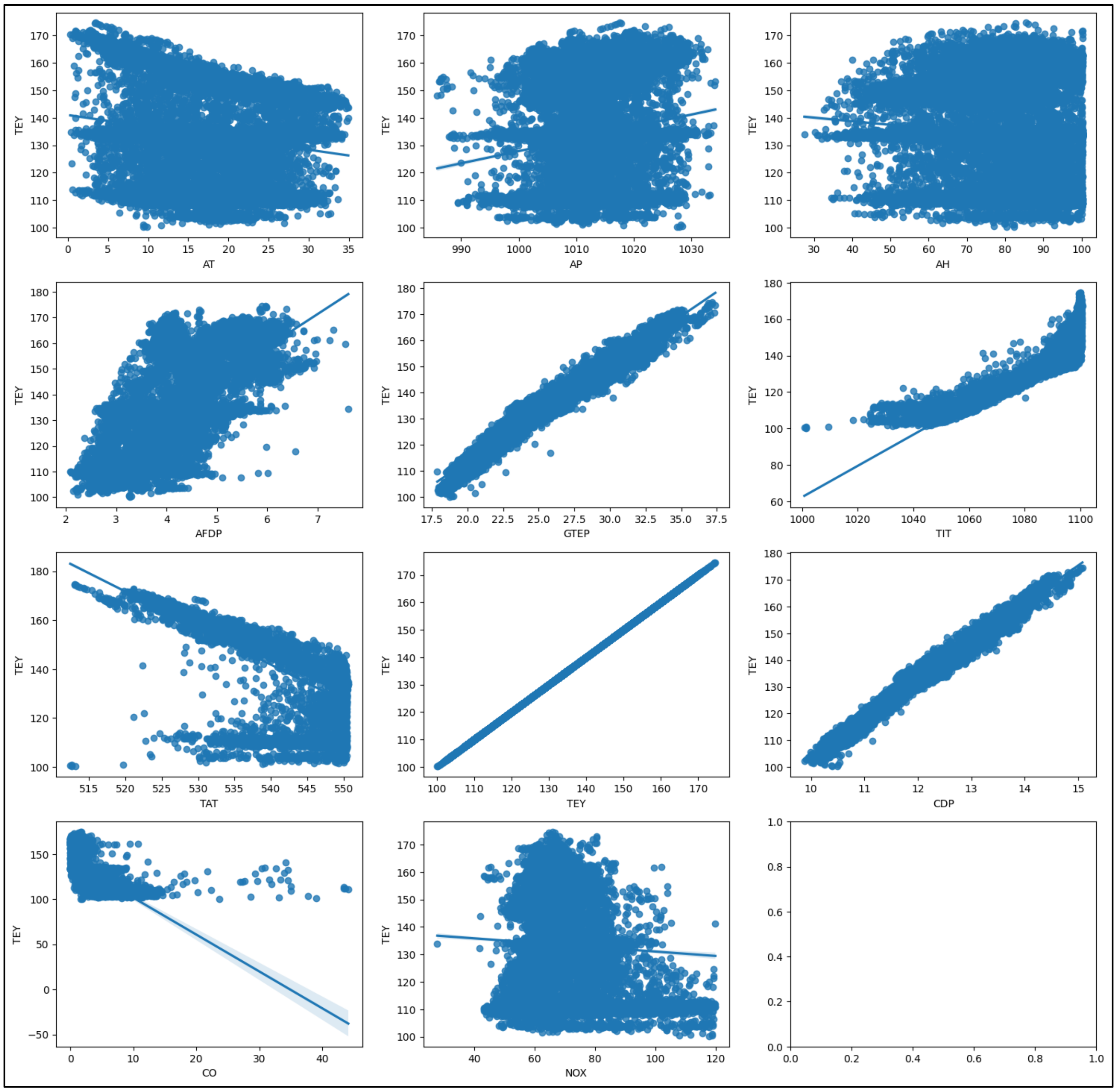
Probability Plots





We can observe that the ambient pressure there follows a normal distribution, the amount of NOx emissions has a normal distribution around the average values and skews around the ends.

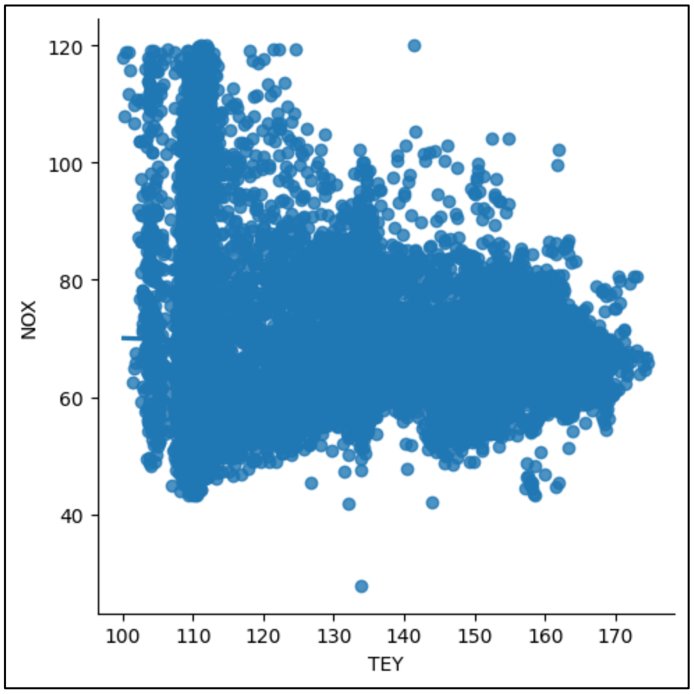
Find the relationship between features and Turbine Energy Yield using Scatter Plots

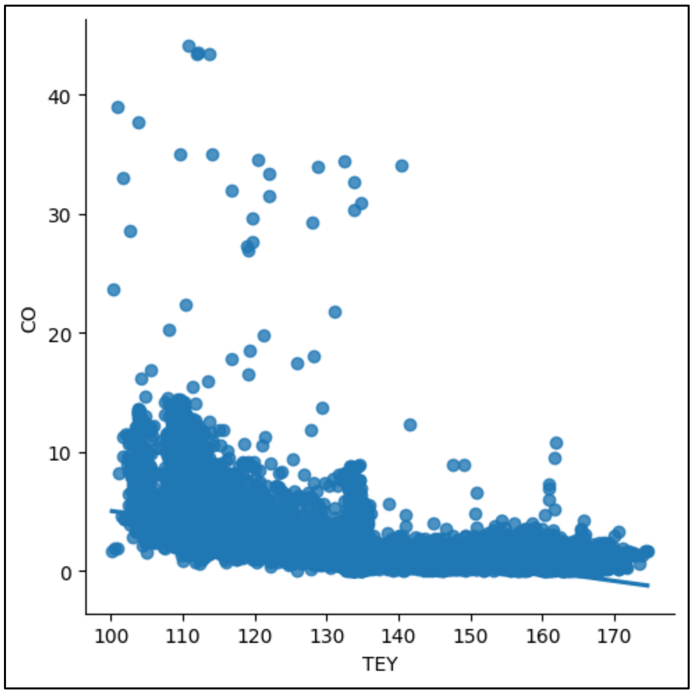


We can double check the relationship by plotting it against TEY, it is observed that GTEP and CDP fall in a straight line, thus follow a similar distribution as the TEY. Thus, we can be confident that GTEP and CDP will be key features on predicting the TEY.

Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca

Gas Emissions and Global Warming Concerns

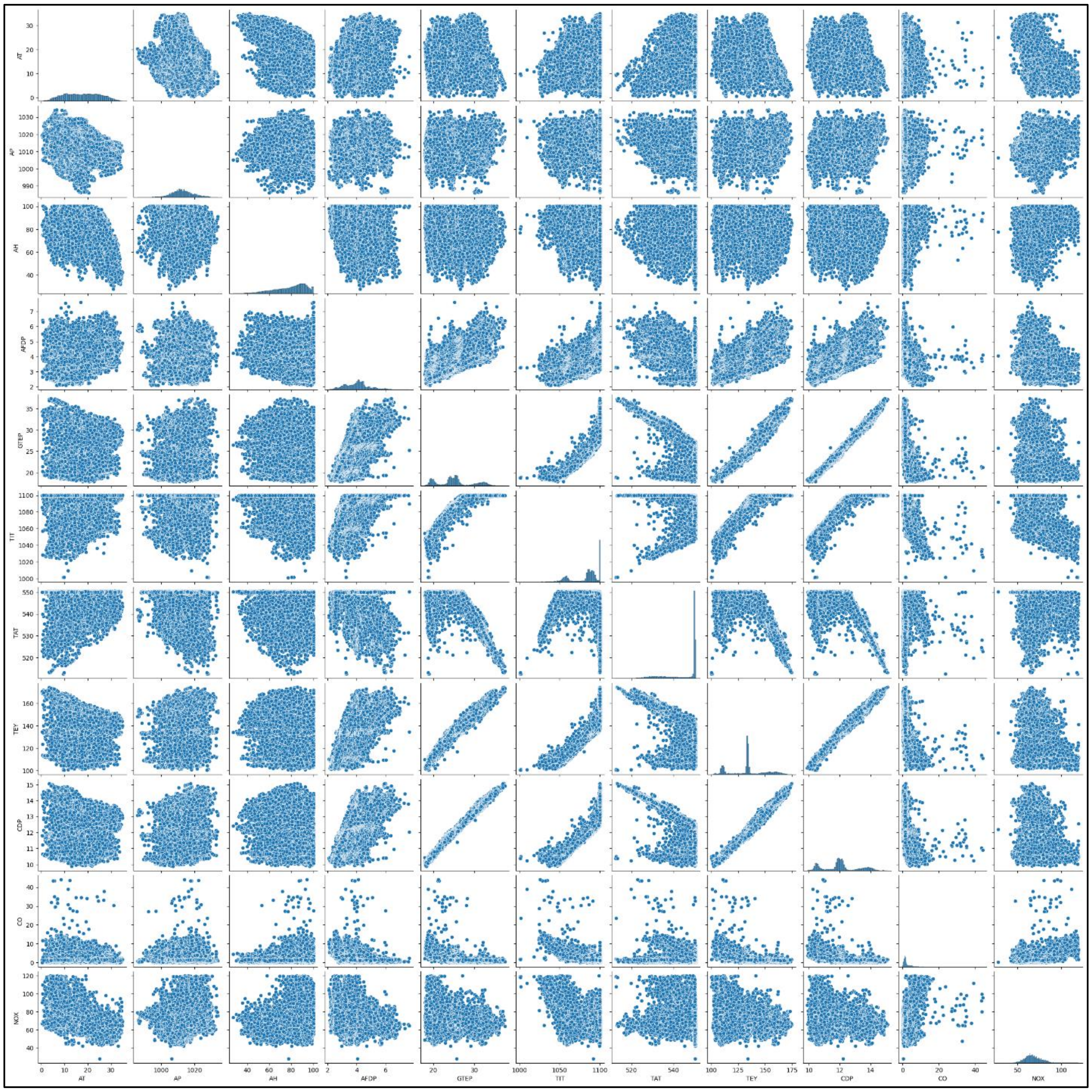




There's also an objective to minimize the emissions of gas turbines while maintaining the purpose of the gas turbine, high yield energy. It seems that a high energy yield does not mean higher CO or NOx emissions

Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca

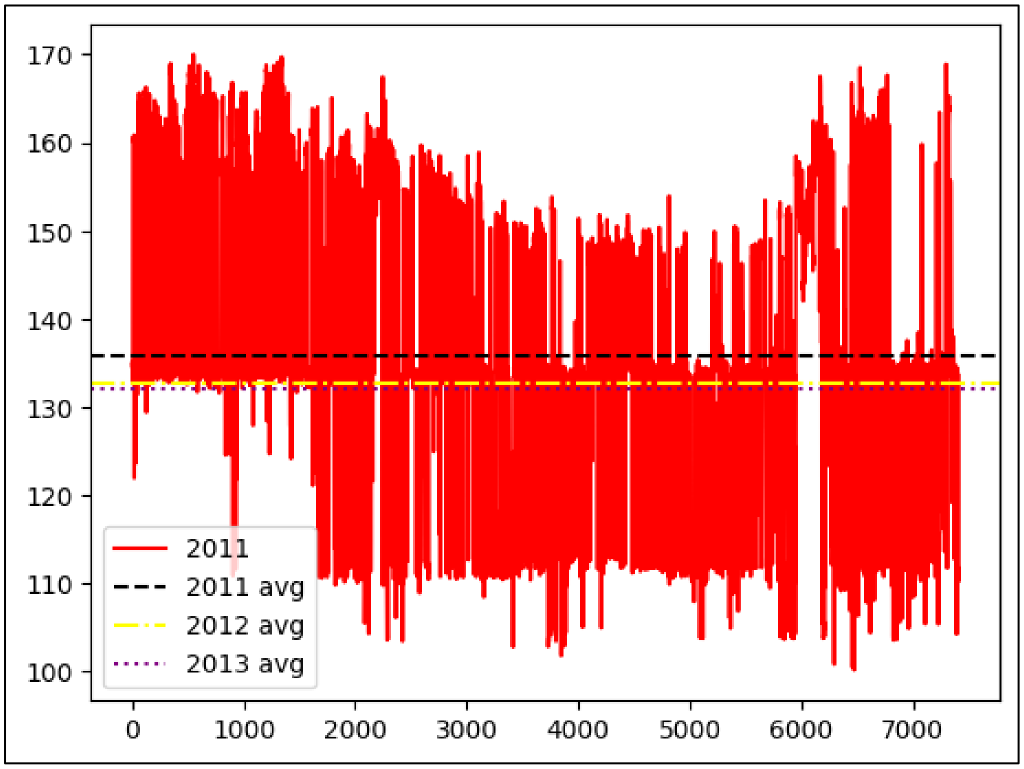
Pairplots – We're looking for graphs that fit a linear relationship with one another

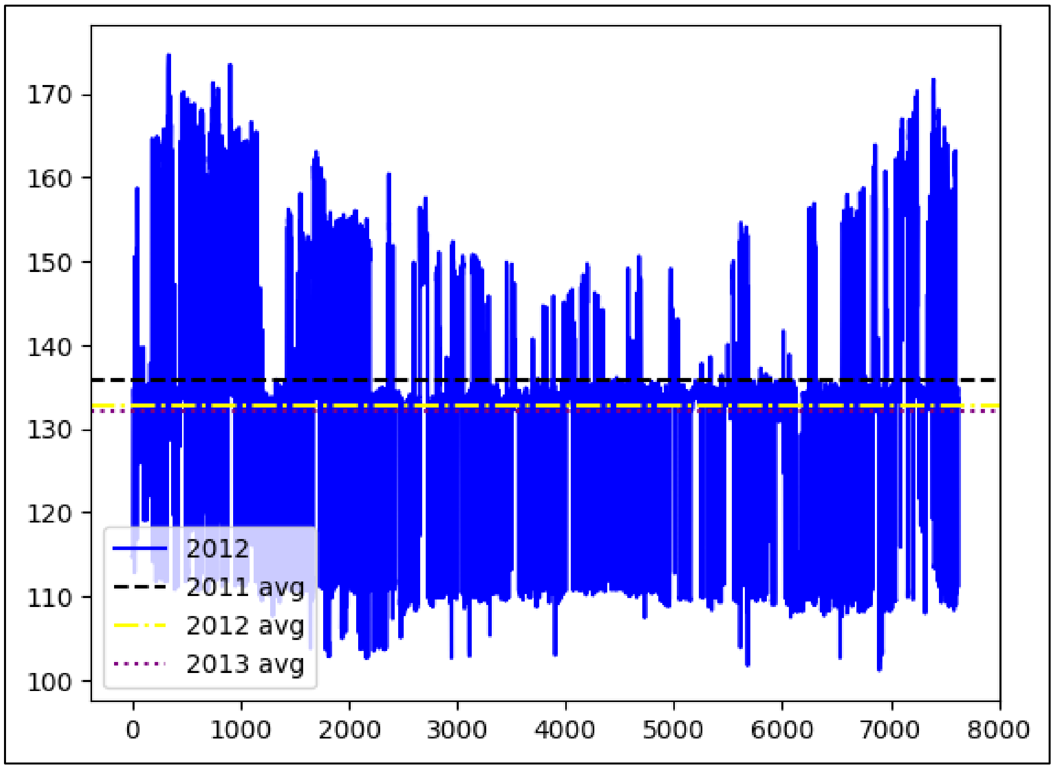


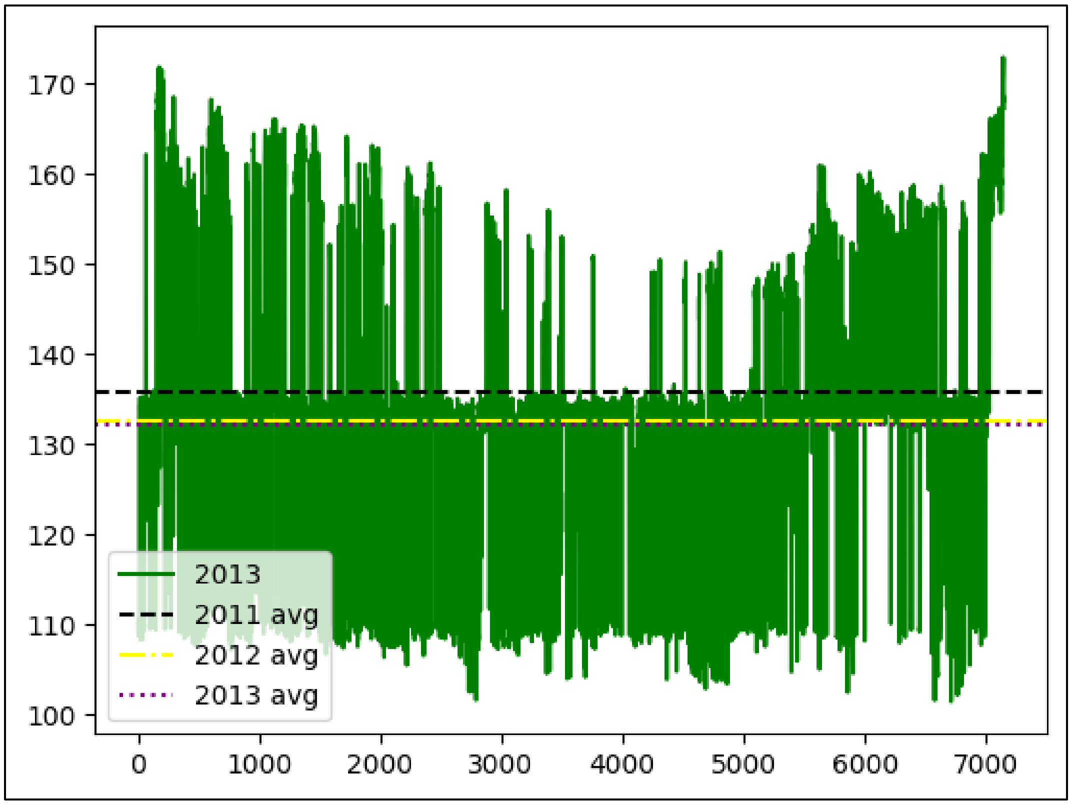
We can see a linear relationship between GTEP and TEY and CDP, so we know those features (GTEP and CDP) will have more feature importance in our machine learning models to predict the TEY.

Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca

Time Series – See if there's any difference between years 2011, 2012, and 2013 since we'll be merging them into 1 dataset as our training data. Each index roughly represents the measurements aggregated over one hour (by means of average or sum) from a gas turbine. Here we compare the TEY over the three years.

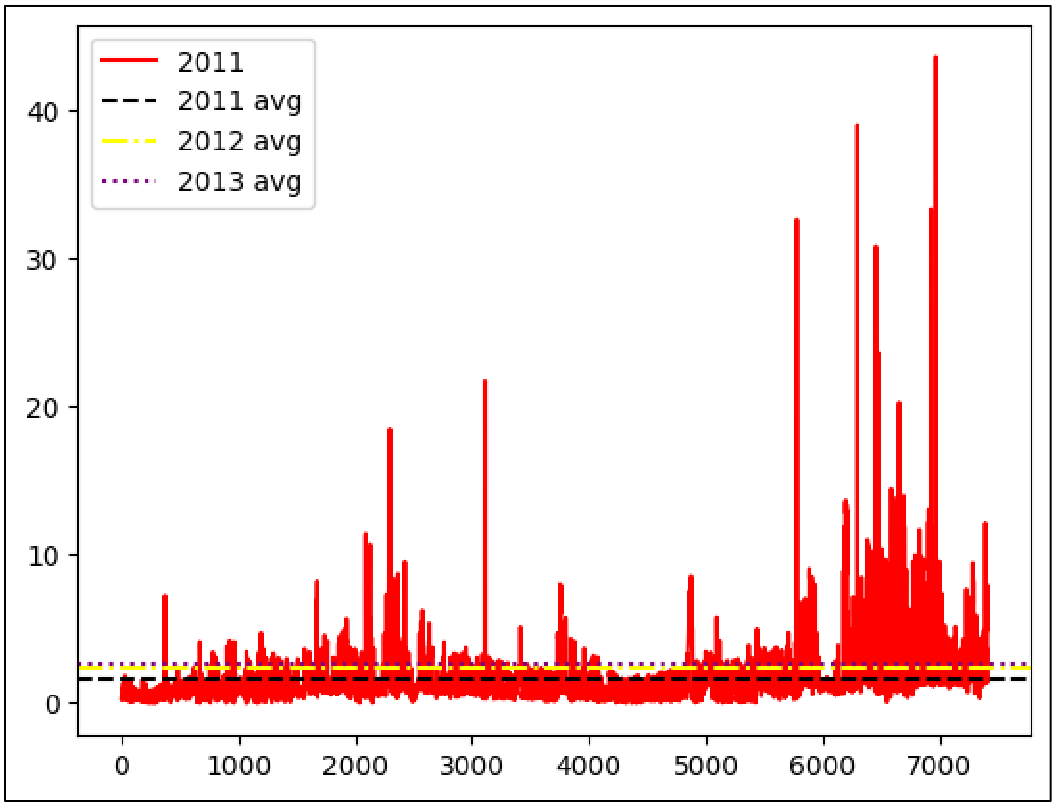


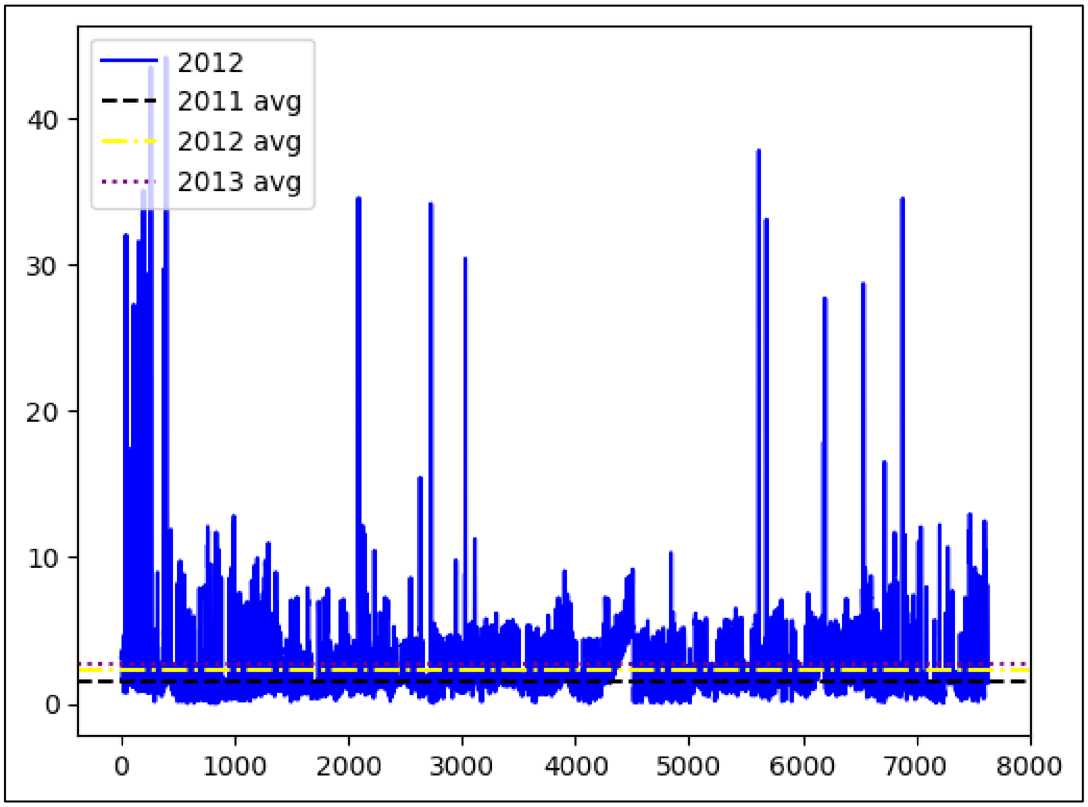




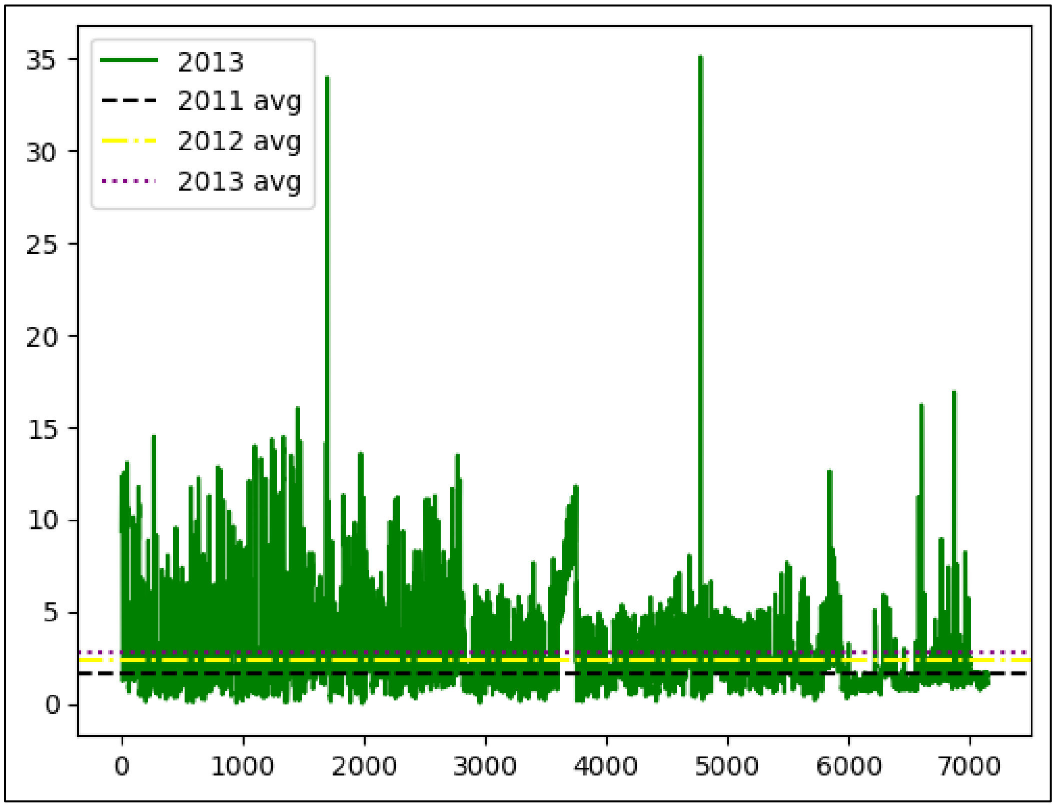
We can see that the minimum and maximum range for the Turbine Energy Yield stays the same throughout the years, we see that the average energy output is slightly decreasing. In 2011, we can observe that the Turbine energy yield was higher for the first quarter, and for the middle of the graphs, there are many more low points, this may be since we require more energy/electricity in the winter compared to the summer.

Compare CO emissions over the three years





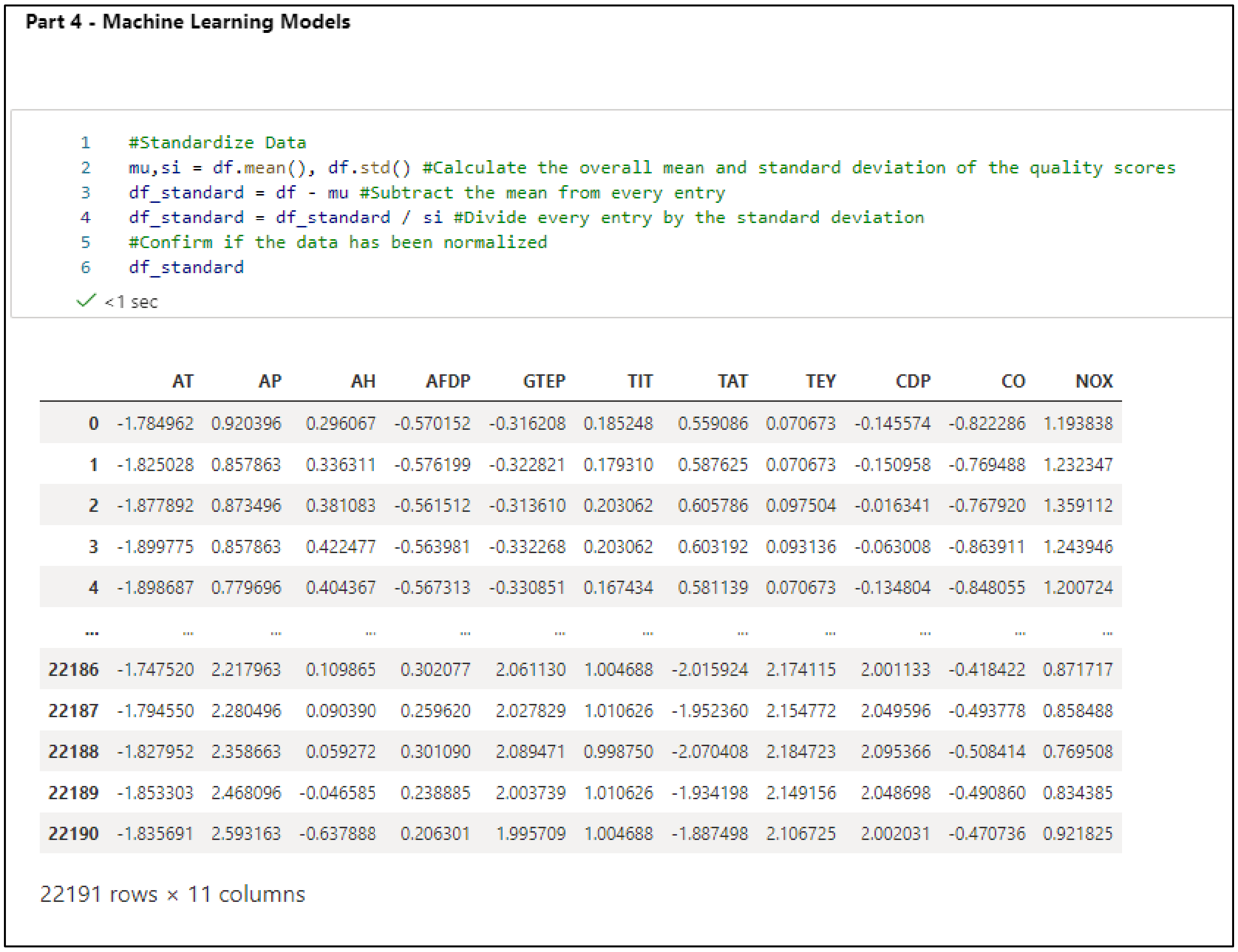
Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca



We can see that the minimum and maximum range for the CO emissions are roughly the same throughout the years, we see that the average CO emissions is increasing over the years. There are a few outliers with high spikes in NOx emissions.

Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca

4. Implement 2 machine learning models and explain which algorithms you have selected and why. Compare them and show success metrics (Accuracy/RMSE/Confusion Matrix) as per your problem. Explain results.



Part 4 - Machine Learning Models

```
#Standardize Data
mu,si = df.mean(), df.std() #Calculate the overall mean and standard deviation of the quality scores
df_standard = df - mu #Subtract the mean from every entry
df_standard = df_standard / si #Divide every entry by the standard deviation
#Confirm if the data has been normalized
df_standard
```

✓ <1 sec

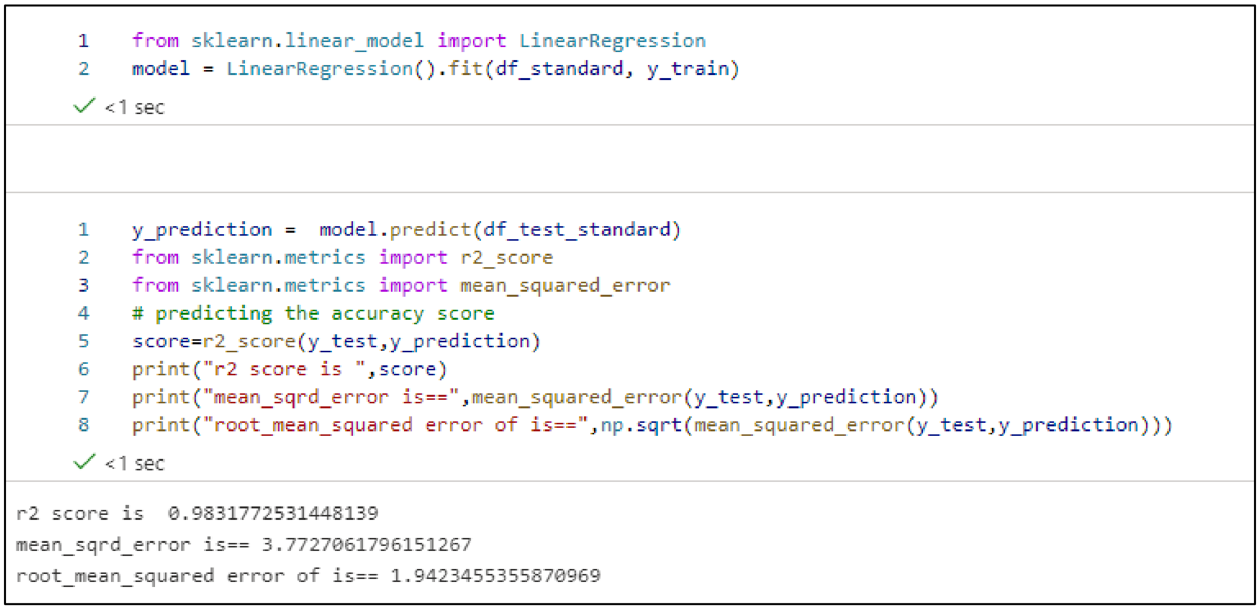
	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
0	-1.784962	0.920396	0.296067	-0.570152	-0.316208	0.185248	0.559086	0.070673	-0.145574	-0.822286	1.193838
1	-1.825028	0.857863	0.336311	-0.576199	-0.322821	0.179310	0.587625	0.070673	-0.150958	-0.769488	1.232347
2	-1.877892	0.873496	0.381083	-0.561512	-0.313610	0.203062	0.605786	0.097504	-0.016341	-0.767920	1.359112
3	-1.899775	0.857863	0.422477	-0.563981	-0.332268	0.203062	0.603192	0.093136	-0.063008	-0.863911	1.243946
4	-1.898687	0.779696	0.404367	-0.567313	-0.330851	0.167434	0.581139	0.070673	-0.134804	-0.848055	1.200724
...	...	...	...	...	...	...	...	...	...	...	...
22186	-1.747520	2.217963	0.109865	0.302077	2.061130	1.004688	-2.015924	2.174115	2.001133	-0.418422	0.871717
22187	-1.794550	2.280496	0.090390	0.259620	2.027829	1.010626	-1.952360	2.154772	2.049596	-0.493778	0.858488
22188	-1.827952	2.358663	0.059272	0.301090	2.089471	0.998750	-2.070408	2.184723	2.095366	-0.508414	0.769508
22189	-1.853303	2.468096	-0.046585	0.238885	2.003739	1.010626	-1.934198	2.149156	2.048698	-0.490860	0.834385
22190	-1.835691	2.593163	-0.637888	0.206301	1.995709	1.004688	-1.887498	2.106725	2.002031	-0.470736	0.921825

22191 rows × 11 columns

Standardized Training data and remove the target column (TEY)

Multiple Linear Regression Model

Multiple Linear Regression model was chosen to predict the Turbine energy yield, this is one of the simplest predictive regression models. We can see if a linear scaling of the features can be summed up to determine the turbine yield energy. Linear regression predicts the weights of each feature. It used the method of ordinary least squares, minimize the sum of squared residuals (SSR) to determine the best weights.



```python
from sklearn.linear_model import LinearRegression
model = LinearRegression().fit(df_standard, y_train)
```

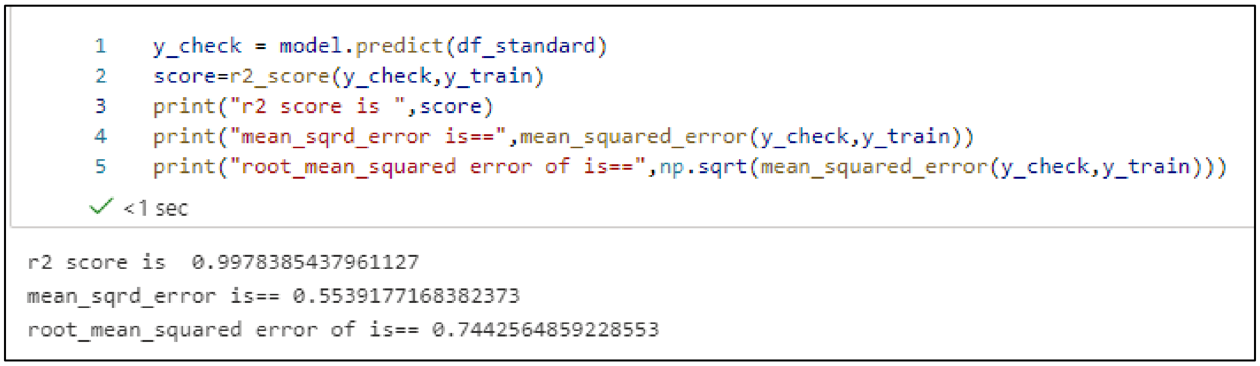
✓ <1 sec

```python
y_prediction =  model.predict(df_test_standard)
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
# predicting the accuracy score
score=r2_score(y_test,y_prediction)
print("r2 score is ",score)
print("mean_sqrd_error is==",mean_squared_error(y_test,y_prediction))
print("root_mean_squared error of is==",np.sqrt(mean_squared_error(y_test,y_prediction)))
```

✓ <1 sec

```
r2 score is  0.9831772531448139
mean_sqrd_error is== 3.7727061796151267
root_mean_squared error of is== 1.9423455355870969
```

The accuracy score is high at about 98.31%, the RMSE is also low at about 1.94 MWH considering its goes on average to 130 MWH. R2 score is the coefficient of determination, the amount of variation in the target (Turbine energy yield) that can be explained by the dependence of the input features. We can further improve the accuracy by hyper tuning the model. But this would be easier to leave the hyper parameter tuning to AutoML in Part B-5.

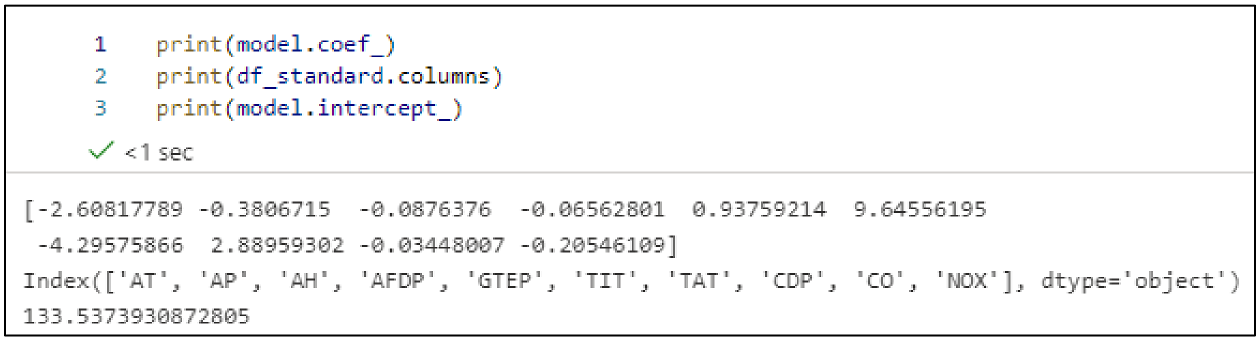


```python
y_check = model.predict(df_standard)
score=r2_score(y_check,y_train)
print("r2 score is ",score)
print("mean_sqrd_error is==",mean_squared_error(y_check,y_train))
print("root_mean_squared error of is==",np.sqrt(mean_squared_error(y_check,y_train)))
```

✓ <1 sec

```
r2 score is  0.9978385437961127
mean_sqrd_error is== 0.5539177168382373
root_mean_squared error of is== 0.7442564859228553
```

We ran the model against the training data again to see whether it was overfitting or underfitting, the model did slightly better on the training data compared to the test data so the model but still has relatively high R2 scores. The simple multi linear regression model is slightly overfitting the training data.



```python
print(model.coef_)
print(df_standard.columns)
print(model.intercept_)
```

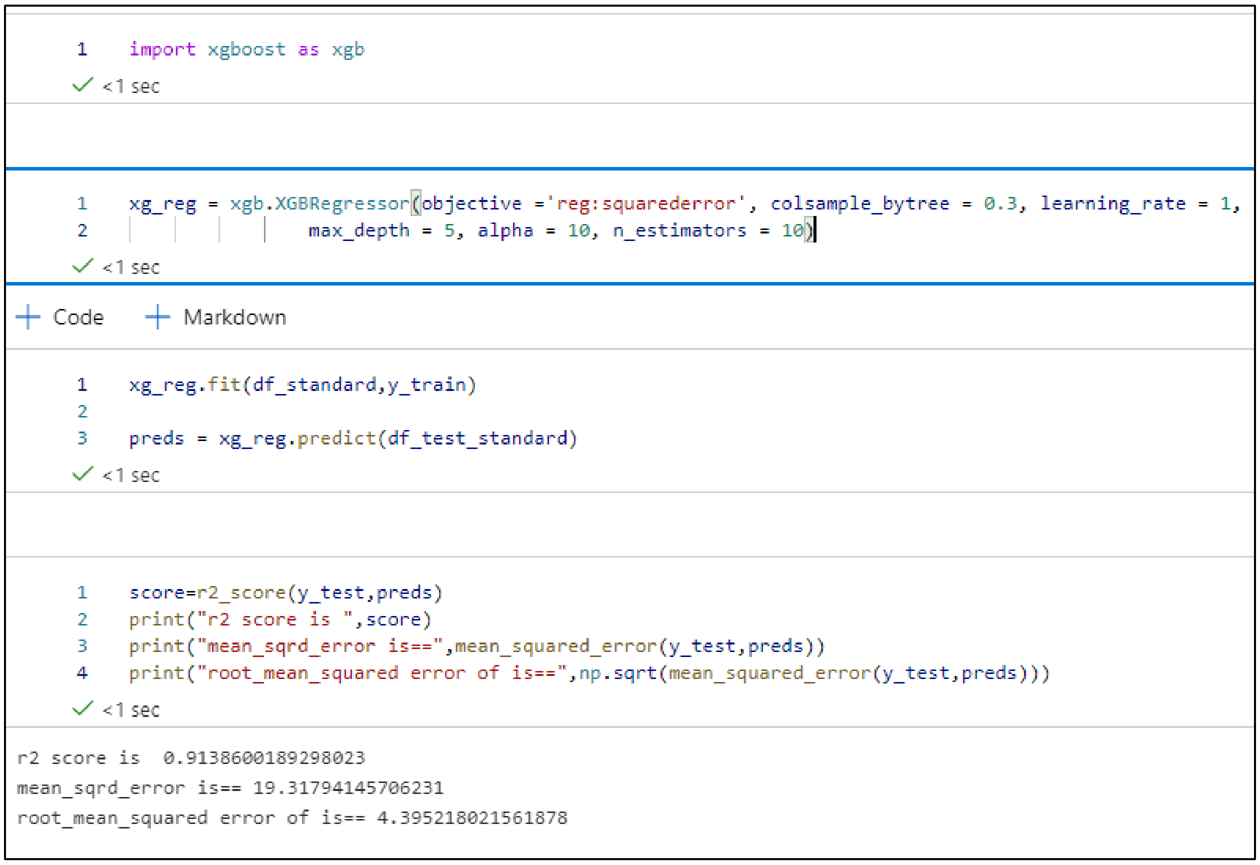
✓ <1 sec

```
[-2.60817789 -0.3806715  -0.0876376  -0.06562801  0.93759214  9.64556195
 -4.29575866  2.88959302 -0.03448007 -0.20546109]
Index(['AT', 'AP', 'AH', 'AFDP', 'GTEP', 'TIT', 'TAT', 'CDP', 'CO', 'NOX'], dtype='object')
133.5373930872805
```

Since the data is normalized, the bias is set at around the average value. We can examine the weight of each feature. We can observe that a higher Turbine inlet temperature and Compressor Discharge pressure is correlated for a higher turbine energy yield. We can also see that Turbine after temperature is negatively correlated with a higher energy yield (most of the energy turning into heat possibly). \

XGBoost

XGBoost stands for Extreme Gradient Boost, boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. XGBoost is a library for developing fast and high-performance gradient boosting tree models which can be used for solving regression type problems.



```
import xgboost as xgb
```
✓ <1 sec

```
xg_reg = xgb.XGBRegressor(objective ='reg:squarederror', colsample_bytree = 0.3, learning_rate = 1,
                max_depth = 5, alpha = 10, n_estimators = 10)
```
✓ <1 sec

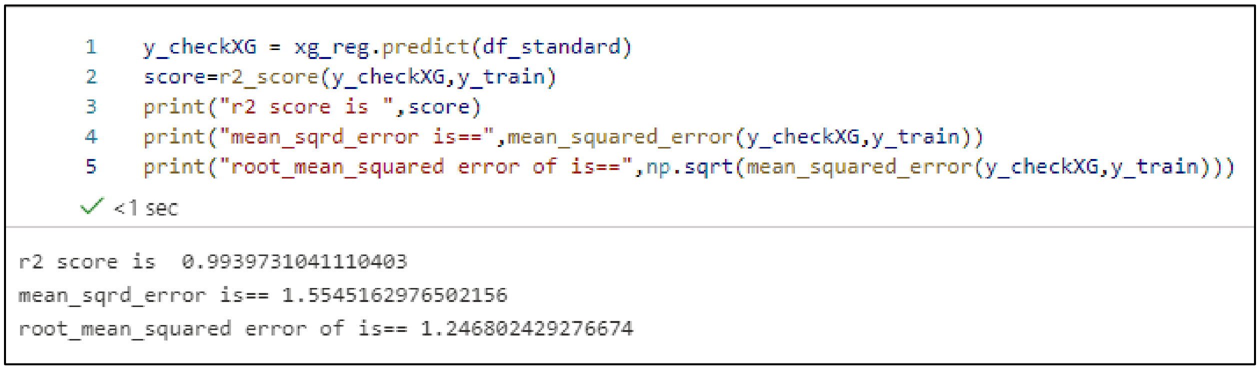
```
xg_reg.fit(df_standard,y_train)

preds = xg_reg.predict(df_test_standard)
```
✓ <1 sec

```
score=r2_score(y_test,preds)
print("r2 score is ",score)
print("mean_sqrd_error is==",mean_squared_error(y_test,preds))
print("root_mean_squared error of is==",np.sqrt(mean_squared_error(y_test,preds)))
```
✓ <1 sec

```
r2 score is  0.9138600189298023
mean_sqrd_error is== 19.31794145706231
root_mean_squared error of is== 4.395218021561878
```

First tried this model with the learning rate as 0.1, and got the RMSE at around 46, this meant that the model was overfitting. Now the RMSE is around 4.395 MWH which is relatively low with a R2 score of 91.3% without too many hyperparameter tuning.



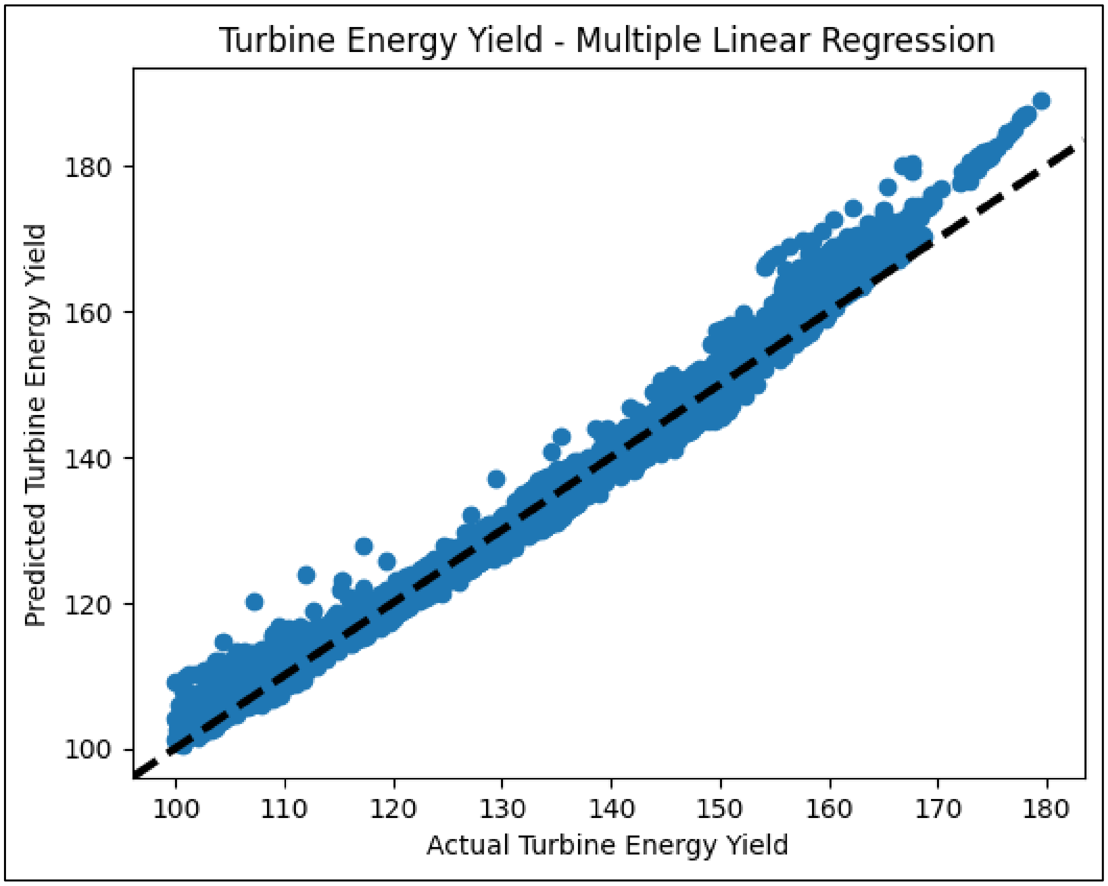
```
y_checkXG = xg_reg.predict(df_standard)
score=r2_score(y_checkXG,y_train)
print("r2 score is ",score)
print("mean_sqrd_error is==",mean_squared_error(y_checkXG,y_train))
print("root_mean_squared error of is==",np.sqrt(mean_squared_error(y_checkXG,y_train)))
```

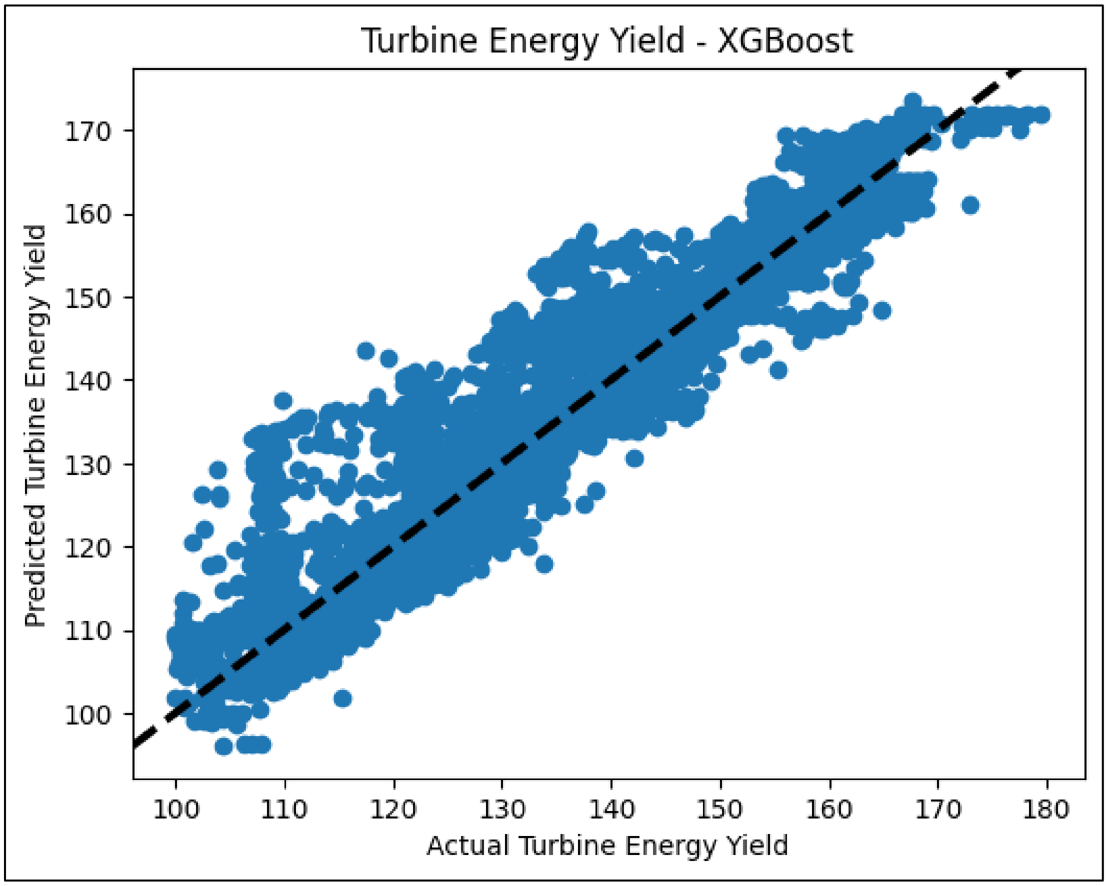
✓ <1 sec

```
r2 score is  0.9939731041110403
mean_sqrd_error is== 1.5545162976502156
root_mean_squared error of is== 1.246802429276674
```

We ran the model against the training data again to see whether it was overfitting or underfitting, the model did slightly better on the training data compared to the test data so the model but still has high R2 scores. The more complex XGBoost model is slightly overfitting the training data.

Comparing the two-machine learning models



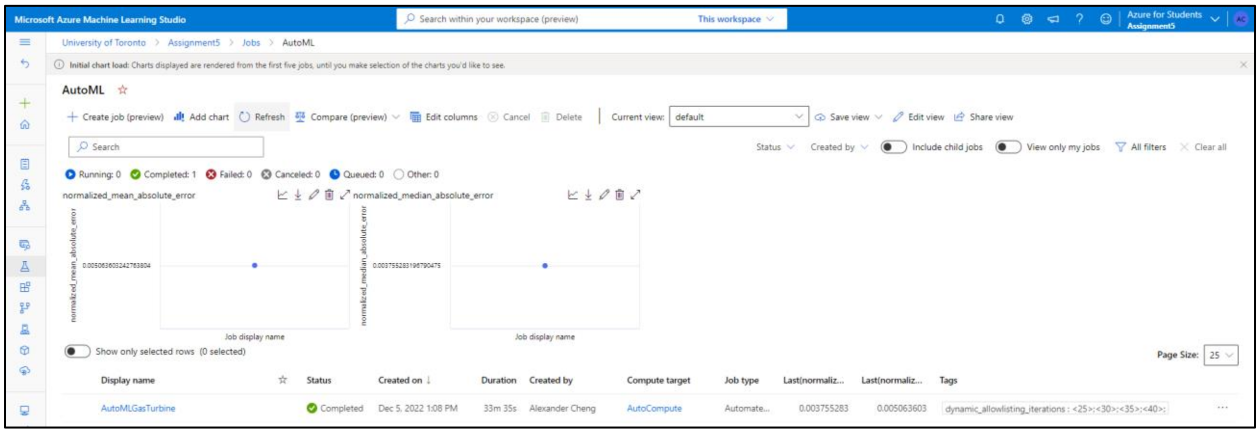


Our simple multiple linear regression model performed better than the complex XGBoost model in terms of having a higher accuracy (R2 score) of (98.31% vs 91.38%) and a lower RMSE value (1.94 vs 4.395 MWH). From the plot, we can see that the multi linear regression model does a good job predicting the values except when it gets to the extremes (100-110 or 160-170 MWH). From the XGBoost model, we can see that we are overfitting our model to the training data since it has such a large spread and relatively low R2 score. If we were to hyper tune the models and run cross fold validations with the training data, the model would give better results. This will be performed in part 5 by the AutoML program.

Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca

5. Deploy a run-time pipeline for your dataset using Azure Designer Studio.
   Or do hyperparameter tuning for your algorithms. Explain your results.
   Or Use Automated ML for your data set. Explain the best model results.

Used Automated ML for the dataset:



Initial chart load: Charts displayed are rendered from the first five jobs, until you make selection of the charts you'd like to see.

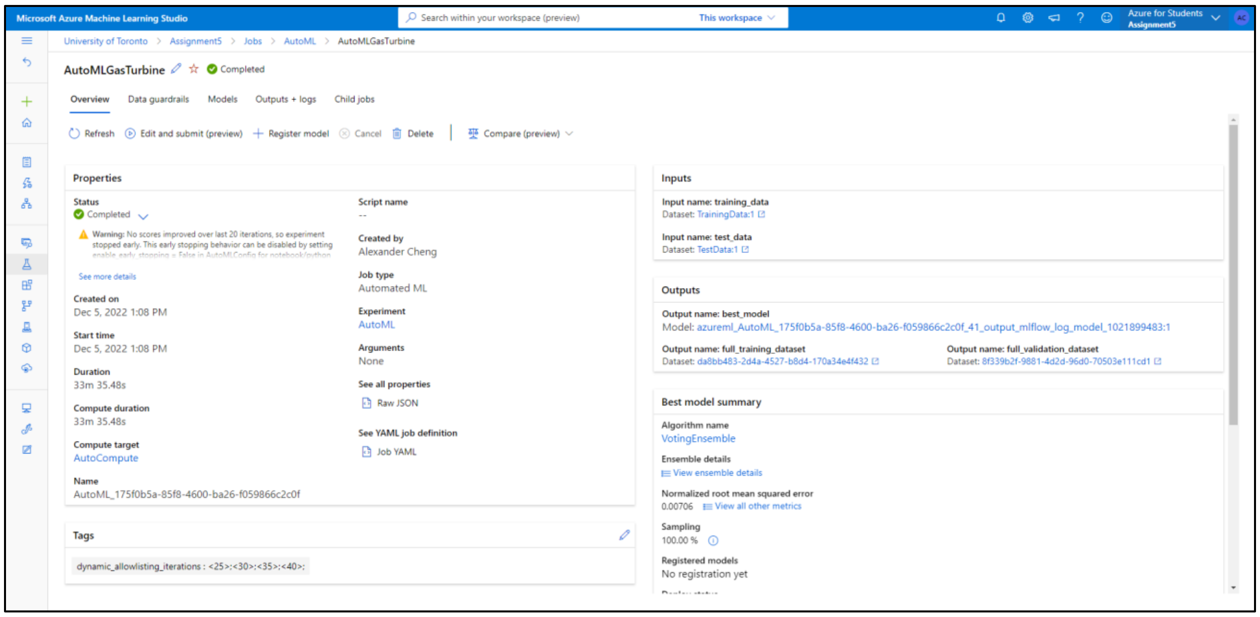
AutoML

Current view: default

Running: 0 Completed: 1 Failed: 0 Canceled: 0 Queued: 0 Other: 0

Display name	Status	Created on	Duration	Created by	Compute target	Job type	Last(normaliz...	Last(normaliz...	Tags
AutoMLGasTurbine	Completed	Dec 5, 2022 1:08 PM	33m 35s	Alexander Cheng	AutoCompute	Automate...	0.003755283	0.005063603	dynamic_allowlisting_iterations : <25>;<30>;<35>;<40>;

AutoML Job for the Gas Turbine Dataset



AutoMLGasTurbine Completed

Properties

Status
Completed

Warning: No scores improved over last 20 iterations, so experiment stopped early. This early stopping behavior can be disabled by setting enable_early_stopping = False in AutoMLConfig for notebook/python

Created on
Dec 5, 2022 1:08 PM

Start time
Dec 5, 2022 1:08 PM

Duration
33m 35.48s

Compute duration
33m 35.48s

Compute target
AutoCompute

Name
AutoML_175f0b5a-85f8-4600-ba26-f059866c2c0f

Script name
--

Created by
Alexander Cheng

Job type
Automated ML

Experiment
AutoML

Arguments
None

See all properties

See YAML job definition

Tags
dynamic_allowlisting_iterations : <25>;<30>;<35>;<40>;

Inputs

Input name: training_data
Dataset: TrainingData:1

Input name: test_data
Dataset: TestData:1

Outputs

Output name: best_model
Model: azureml_AutoML_175f0b5a-85f8-4600-ba26-f059866c2c0f_41_output_mlflow_log_model_1021899483:1

Output name: full_training_dataset
Dataset: da8bb483-2d4a-4527-b8d4-170a34e4f432

Output name: full_validation_dataset
Dataset: 8f339b2f-9881-4d2d-96d0-70503e111cd1

Best model summary

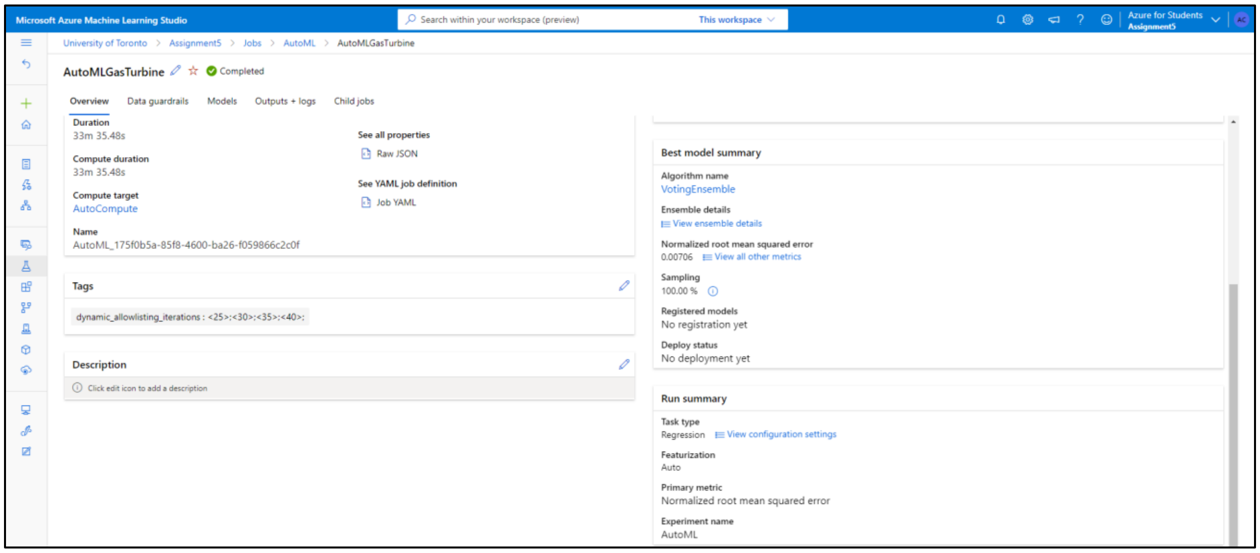
Algorithm name
VotingEnsemble

Ensemble details

Normalized root mean squared error
0.00706

Sampling
100.00 %

Registered models
No registration yet


Microsoft Azure Machine Learning Studio

University of Toronto > Assignment5 > Jobs > AutoML > AutoMLGasTurbine

AutoMLGasTurbine Completed

Overview | Data guardrails | Models | Outputs + logs | Child jobs

Duration
33m 35.48s

Compute duration
33m 35.48s

Compute target
AutoCompute

Name
AutoML_175f0b5a-85f8-4600-ba26-f059866c2c0f

See all properties
Raw JSON

See YAML job definition
Job YAML

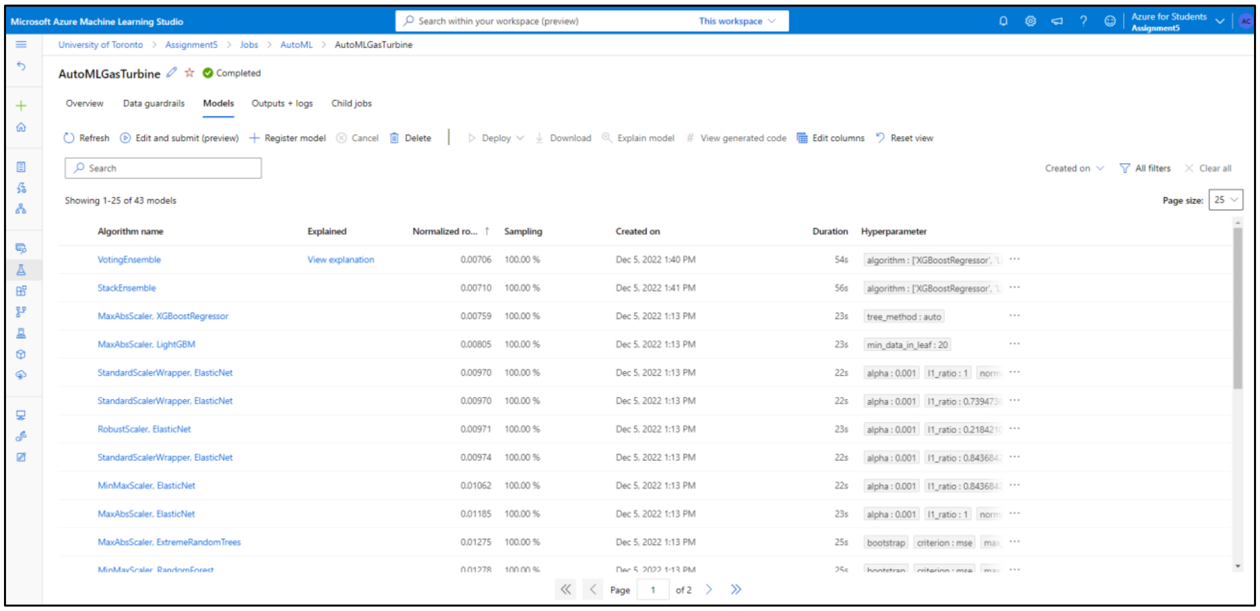
Tags
dynamic_allowlisting_iterations : <25>;<30>;<35>;<40>;

Description
Click edit icon to add a description

Best model summary
Algorithm name
VotingEnsemble
Ensemble details
View ensemble details
Normalized root mean squared error
0.00706 View all other metrics
Sampling
100.00 %
Registered models
No registration yet
Deploy status
No deployment yet

Run summary
Task type
Regression View configuration settings
Featurization
Auto
Primary metric
Normalized root mean squared error
Experiment name
AutoML

List of models that Azure tested to determine which model provides the best results:


Microsoft Azure Machine Learning Studio

University of Toronto > Assignment5 > Jobs > AutoML > AutoMLGasTurbine

AutoMLGasTurbine Completed

Overview | Data guardrails | Models | Outputs + logs | Child jobs

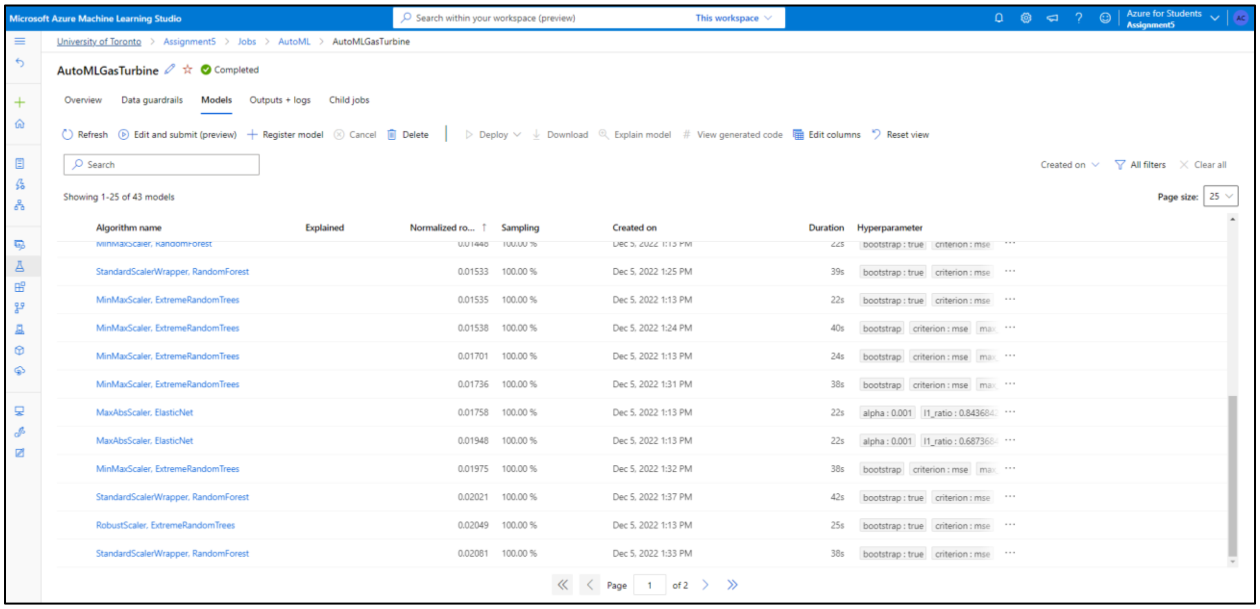
Refresh | Edit and submit (preview) | Register model | Cancel | Delete | Deploy | Download | Explain model | View generated code | Edit columns | Reset view

Search | Created on | All filters | Clear all

Showing 1-25 of 43 models | Page size: 25

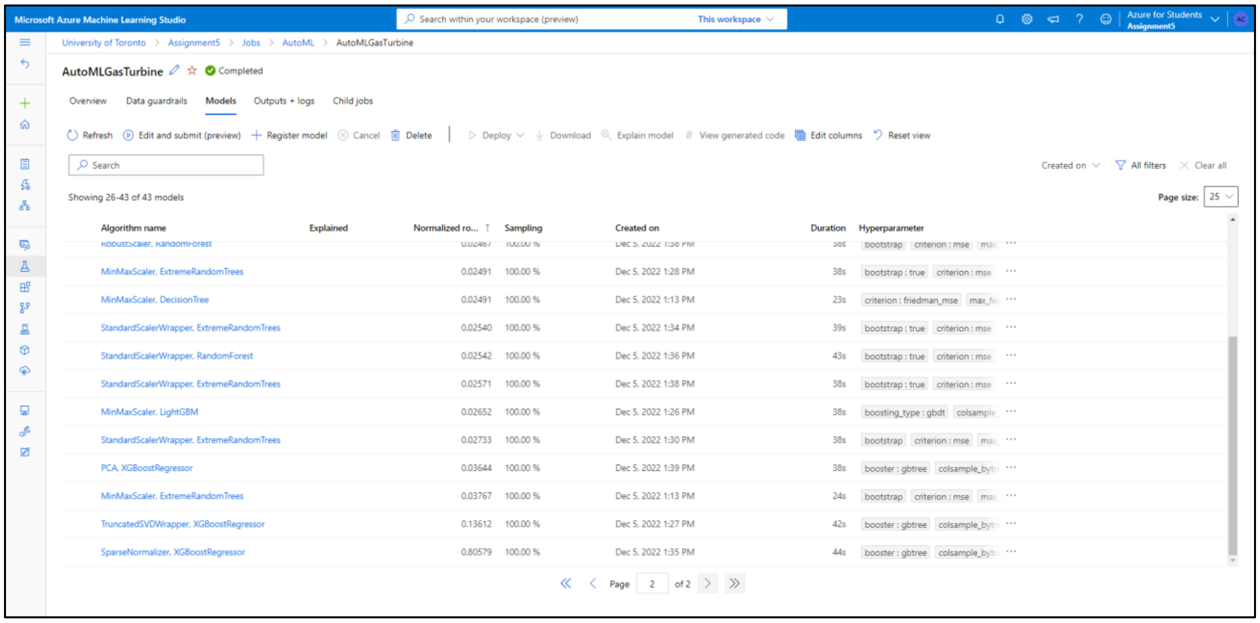
Algorithm name	Explained	Normalized ro... ↑	Sampling	Created on	Duration	Hyperparameter
VotingEnsemble	View explanation	0.00706	100.00 %	Dec 5, 2022 1:40 PM	54s	algorithm : ['XGBoostRegressor', 'L
StackEnsemble		0.00710	100.00 %	Dec 5, 2022 1:41 PM	56s	algorithm : ['XGBoostRegressor', 'L
MaxAbsScaler, XGBoostRegressor		0.00759	100.00 %	Dec 5, 2022 1:13 PM	23s	tree_method : auto
MaxAbsScaler, LightGBM		0.00805	100.00 %	Dec 5, 2022 1:13 PM	23s	min_data_in_leaf : 20
StandardScalerWrapper, ElasticNet		0.00970	100.00 %	Dec 5, 2022 1:13 PM	22s	alpha : 0.001 l1_ratio : 1 norm
StandardScalerWrapper, ElasticNet		0.00970	100.00 %	Dec 5, 2022 1:13 PM	22s	alpha : 0.001 l1_ratio : 0.739473
RobustScaler, ElasticNet		0.00971	100.00 %	Dec 5, 2022 1:13 PM	23s	alpha : 0.001 l1_ratio : 0.218421
StandardScalerWrapper, ElasticNet		0.00974	100.00 %	Dec 5, 2022 1:13 PM	22s	alpha : 0.001 l1_ratio : 0.843684
MinMaxScaler, ElasticNet		0.01062	100.00 %	Dec 5, 2022 1:13 PM	22s	alpha : 0.001 l1_ratio : 0.843684
MaxAbsScaler, ElasticNet		0.01185	100.00 %	Dec 5, 2022 1:13 PM	23s	alpha : 0.001 l1_ratio : 1 norm
MaxAbsScaler, ExtremeRandomTrees		0.01275	100.00 %	Dec 5, 2022 1:13 PM	25s	bootstrap criterion : mse max
MinMaxScaler, RandomForest		0.01278	100.00 %	Dec 5, 2022 1:13 PM	25s	bootstrap criterion : mse max

Page 1 of 2



Showing 1-25 of 43 models

Algorithm name	Explained	Normalized ro... ↑	Sampling	Created on	Duration	Hyperparameter
MinMaxScaler, RandomForest		0.01448	100.00 %	Dec 5, 2022 1:13 PM	22s	bootstrap : true · criterion : mse ···
StandardScalerWrapper, RandomForest		0.01533	100.00 %	Dec 5, 2022 1:25 PM	39s	bootstrap : true · criterion : mse ···
MinMaxScaler, ExtremeRandomTrees		0.01535	100.00 %	Dec 5, 2022 1:13 PM	22s	bootstrap : true · criterion : mse ···
MinMaxScaler, ExtremeRandomTrees		0.01538	100.00 %	Dec 5, 2022 1:24 PM	40s	bootstrap · criterion : mse · max ···
MinMaxScaler, ExtremeRandomTrees		0.01701	100.00 %	Dec 5, 2022 1:13 PM	24s	bootstrap · criterion : mse · max ···
MinMaxScaler, ExtremeRandomTrees		0.01736	100.00 %	Dec 5, 2022 1:31 PM	38s	bootstrap · criterion : mse · max ···
MaxAbsScaler, ElasticNet		0.01758	100.00 %	Dec 5, 2022 1:13 PM	22s	alpha : 0.001 · l1_ratio : 0.843684 ···
MaxAbsScaler, ElasticNet		0.01948	100.00 %	Dec 5, 2022 1:13 PM	22s	alpha : 0.001 · l1_ratio : 0.687368 ···
MinMaxScaler, ExtremeRandomTrees		0.01975	100.00 %	Dec 5, 2022 1:32 PM	38s	bootstrap · criterion : mse · max ···
StandardScalerWrapper, RandomForest		0.02021	100.00 %	Dec 5, 2022 1:37 PM	42s	bootstrap : true · criterion : mse ···
RobustScaler, ExtremeRandomTrees		0.02049	100.00 %	Dec 5, 2022 1:13 PM	25s	bootstrap : true · criterion : mse ···
StandardScalerWrapper, RandomForest		0.02081	100.00 %	Dec 5, 2022 1:33 PM	38s	bootstrap : true · criterion : mse ···

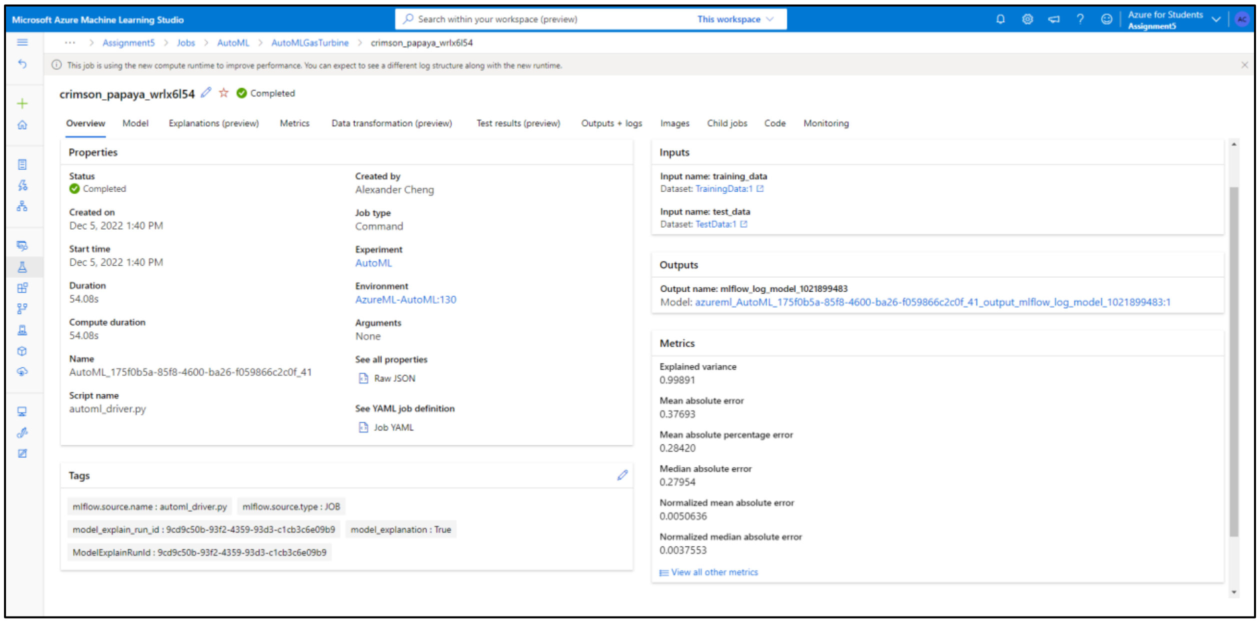


Showing 26-43 of 43 models

Algorithm name	Explained	Normalized ro... ↑	Sampling	Created on	Duration	Hyperparameter
RobustScaler, RandomForest		0.02467	100.00 %	Dec 5, 2022 1:38 PM	38s	bootstrap · criterion : mse · max ···
MinMaxScaler, ExtremeRandomTrees		0.02491	100.00 %	Dec 5, 2022 1:28 PM	38s	bootstrap : true · criterion : mse ···
MinMaxScaler, DecisionTree		0.02491	100.00 %	Dec 5, 2022 1:13 PM	23s	criterion : friedman_mse · max_fe ···
StandardScalerWrapper, ExtremeRandomTrees		0.02540	100.00 %	Dec 5, 2022 1:34 PM	39s	bootstrap : true · criterion : mse ···
StandardScalerWrapper, RandomForest		0.02542	100.00 %	Dec 5, 2022 1:36 PM	43s	bootstrap : true · criterion : mse ···
StandardScalerWrapper, ExtremeRandomTrees		0.02571	100.00 %	Dec 5, 2022 1:38 PM	38s	bootstrap : true · criterion : mse ···
MinMaxScaler, LightGBM		0.02652	100.00 %	Dec 5, 2022 1:26 PM	38s	boosting_type : gbdt · colsample ···
StandardScalerWrapper, ExtremeRandomTrees		0.02733	100.00 %	Dec 5, 2022 1:30 PM	38s	bootstrap · criterion : mse · max ···
PCA, XGBoostRegressor		0.03644	100.00 %	Dec 5, 2022 1:39 PM	38s	booster : gbtree · colsample_bytr ···
MinMaxScaler, ExtremeRandomTrees		0.03767	100.00 %	Dec 5, 2022 1:13 PM	24s	bootstrap · criterion : mse · max ···
TruncatedSVDWrapper, XGBoostRegressor		0.13612	100.00 %	Dec 5, 2022 1:27 PM	42s	booster : gbtree · colsample_bytr ···
SparseNormalizer, XGBoostRegressor		0.80579	100.00 %	Dec 5, 2022 1:35 PM	44s	booster : gbtree · colsample_bytr ···

Out of the multiple models that were tested, Voting Ensemble was the best model chosen. A voting ensemble (or a "majority voting ensemble") is an ensemble machine learning model that combines the predictions from multiple other models. It is a technique that may be used to improve model performance, ideally achieving better performance than any single model used in the ensemble. We can further investigate the best model.

Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca



Microsoft Azure Machine Learning Studio

··· > Assignment5 > Jobs > AutoML > AutoMLGasTurbine > crimson_papaya_wrlx6l54

This job is using the new compute runtime to improve performance. You can expect to see a different log structure along with the new runtime.

crimson_papaya_wrlx6l54 Completed

Overview | Model | Explanations (preview) | Metrics | Data transformation (preview) | Test results (preview) | Outputs + logs | Images | Child jobs | Code | Monitoring

Properties

Status: Completed	Created by: Alexander Cheng
Created on: Dec 5, 2022 1:40 PM	Job type: Command
Start time: Dec 5, 2022 1:40 PM	Experiment: AutoML
Duration: 54.08s	Environment: AzureML-AutoML:130
Compute duration: 54.08s	Arguments: None
Name: AutoML_175f0b5a-85f8-4600-ba26-f059866c2c0f_41	See all properties: Raw JSON
Script name: automl_driver.py	See YAML job definition: Job YAML

Tags

mlflow.source.name : automl_driver.py
mlflow.source.type : JOB
model_explain_run_id : 9cd9c50b-93f2-4359-93d3-c1cb3c6e09b9
model_explanation : True
ModelExplainRunId : 9cd9c50b-93f2-4359-93d3-c1cb3c6e09b9

Inputs

Input name: training_data
Dataset: TrainingData:1

Input name: test_data
Dataset: TestData:1

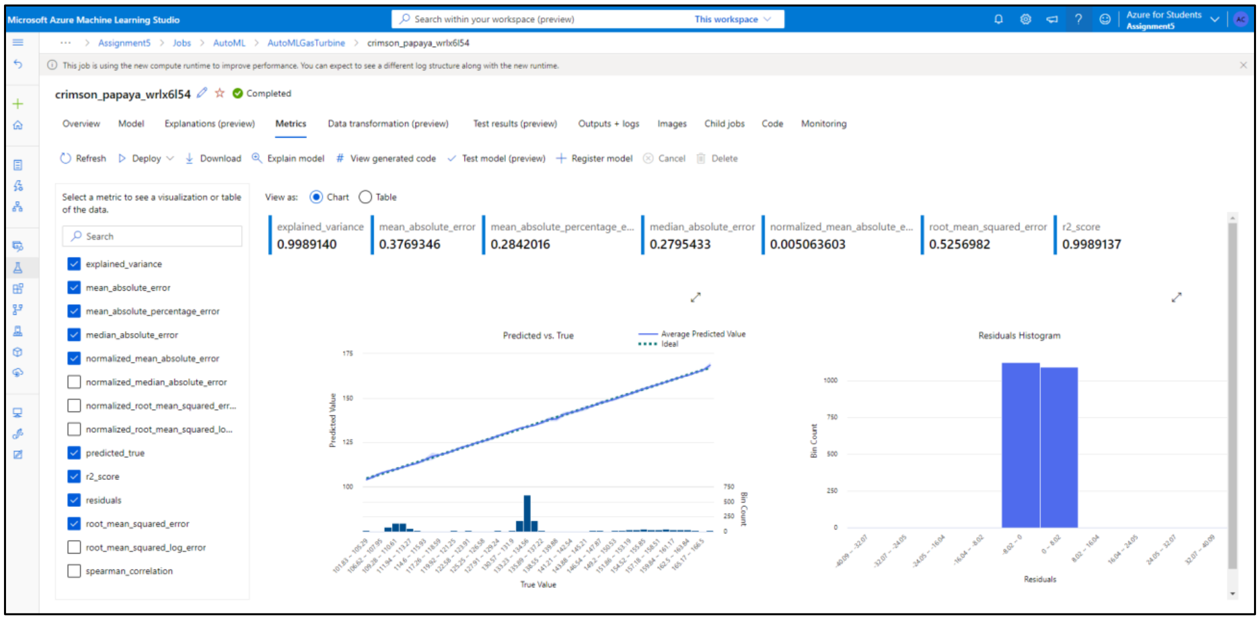
Outputs

Output name: mlflow_log_model_1021899483
Model: azureml_AutoML_175f0b5a-85f8-4600-ba26-f059866c2c0f_41_output_mlflow_log_model_1021899483:1

Metrics

Explained variance	0.99891
Mean absolute error	0.37693
Mean absolute percentage error	0.28420
Median absolute error	0.27954
Normalized mean absolute error	0.0050636
Normalized median absolute error	0.0037553

View all other metrics



Microsoft Azure Machine Learning Studio

··· > Assignment5 > Jobs > AutoML > AutoMLGasTurbine > crimson_papaya_wrlx6l54

This job is using the new compute runtime to improve performance. You can expect to see a different log structure along with the new runtime.

crimson_papaya_wrlx6l54 Completed

Overview | Model | Explanations (preview) | Metrics | Data transformation (preview) | Test results (preview) | Outputs + logs | Images | Child jobs | Code | Monitoring

Refresh | Deploy | Download | Explain model | View generated code | Test model (preview) | Register model | Cancel | Delete

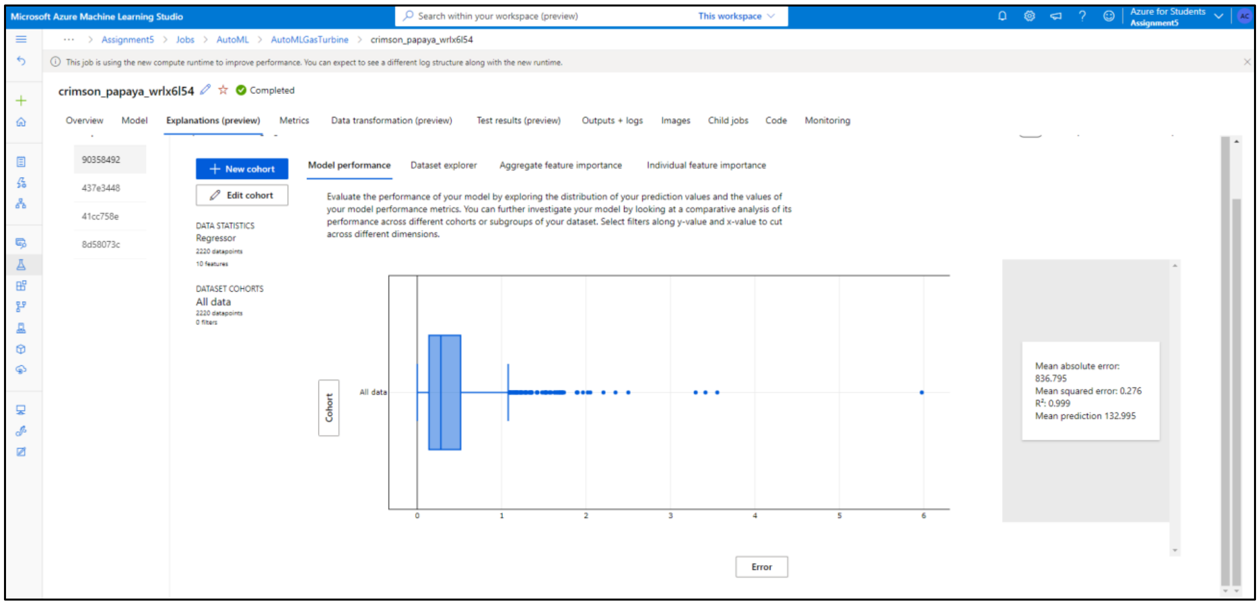
Select a metric to see a visualization or table of the data.

- [✓] explained_variance
- [✓] mean_absolute_error
- [✓] mean_absolute_percentage_error
- [✓] median_absolute_error
- [✓] normalized_mean_absolute_error
- [ ] normalized_median_absolute_error
- [ ] normalized_root_mean_squared_err...
- [ ] normalized_root_mean_squared_lo...
- [✓] predicted_true
- [✓] r2_score
- [✓] residuals
- [✓] root_mean_squared_error
- [ ] root_mean_squared_log_error
- [ ] spearman_correlation

View as: [✓ Chart] Table

explained_variance	mean_absolute_error	mean_absolute_percentage_e...	median_absolute_error	normalized_mean_absolute_e...	root_mean_squared_error	r2_score
0.9989140	0.3769346	0.2842016	0.2795433	0.005063603	0.5256982	0.9989137

It has a very high accuracy (R2 score) and relatively small RMSE value. It can explain most the variance with the features provided.



This job is using the new compute runtime to improve performance. You can expect to see a different log structure along with the new runtime.

crimson_papaya_wrlx6l54 Completed

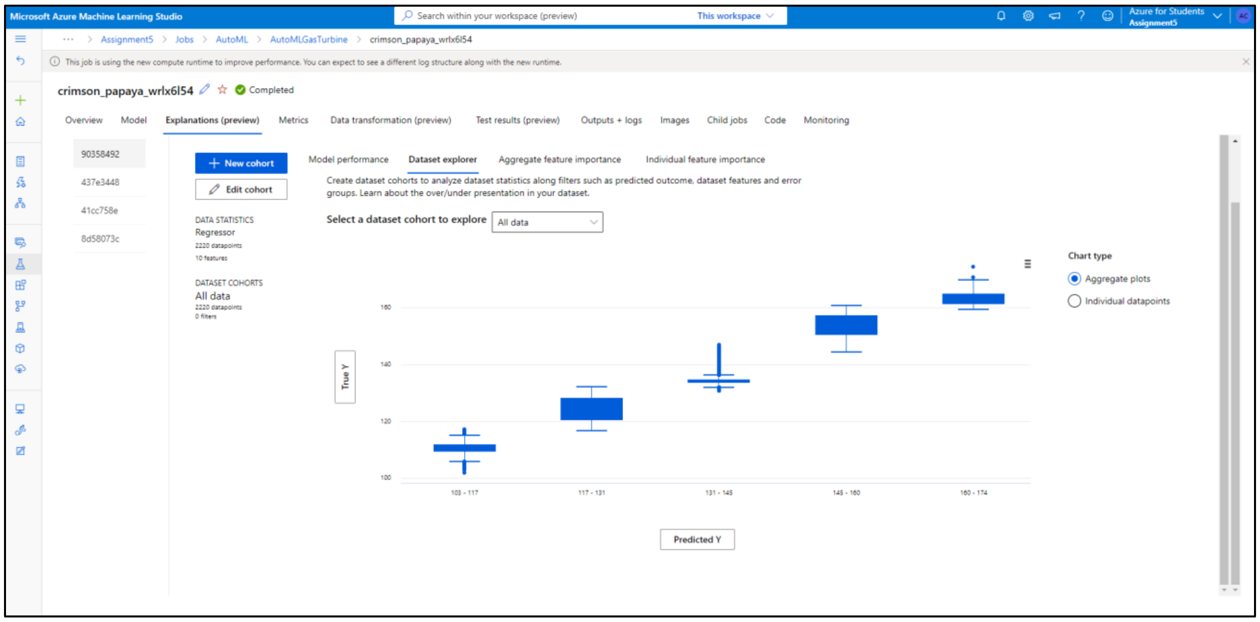
DATA STATISTICS
Regressor
2220 datapoints
10 features

DATASET COHORTS
All data
2220 datapoints
0 filters

Evaluate the performance of your model by exploring the distribution of your prediction values and the values of your model performance metrics. You can further investigate your model by looking at a comparative analysis of its performance across different cohorts or subgroups of your dataset. Select filters along y-value and x-value to cut across different dimensions.

Mean absolute error: 836.795
Mean squared error: 0.276
R^2: 0.999
Mean prediction 132.995

We can see that most of the mean squared error is close to 0.276, RMSE is 0.5256 MWH, and the average prediction of the energy yield is 133 MWH. There are still a few outliers predicted by the model that results in large errors. The R2 score is much higher than the multi linear regression model and the XGboost model I used in Part B-4. The RMSE is also significant smaller as well compares to those two models.



This job is using the new compute runtime to improve performance. You can expect to see a different log structure along with the new runtime.

crimson_papaya_wrlx6l54 Completed

DATA STATISTICS
Regressor
2220 datapoints
10 features

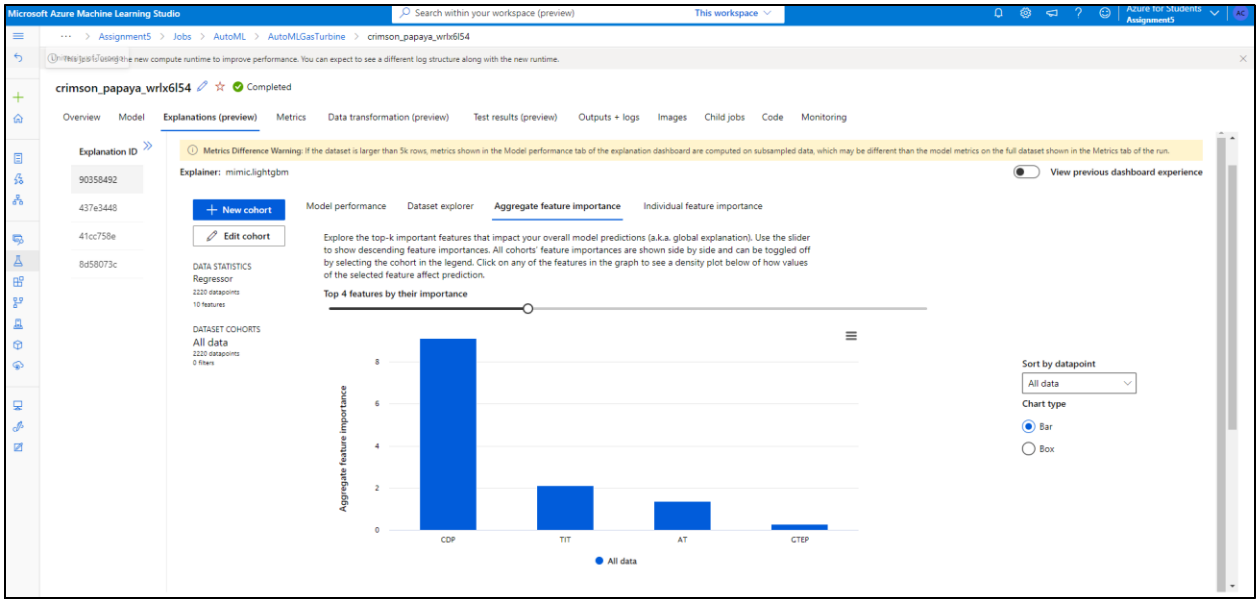
DATASET COHORTS
All data
2220 datapoints
0 filters

Create dataset cohorts to analyze dataset statistics along filters such as predicted outcome, dataset features and error groups. Learn about the over/under presentation in your dataset.

Select a dataset cohort to explore: All data

Chart type: [✓ Aggregate plots] Individual datapoints

Comparing the true and predicted Turbine energy yield, we can observe that the model has the most outliers near the average value, slightly overestimating the energy yield.


Microsoft Azure Machine Learning Studio

Assignment5 > Jobs > AutoML > AutoMLGasTurbine > crimson_papaya_wrlx6l54

This job is using the new compute runtime to improve performance. You can expect to see a different log structure along with the new runtime.

crimson_papaya_wrlx6l54 Completed

Overview | Model | Explanations (preview) | Metrics | Data transformation (preview) | Test results (preview) | Outputs + logs | Images | Child jobs | Code | Monitoring

Explanation ID: 90358492, 437e3448, 41cc758e, 8d58073c

Metrics Difference Warning: If the dataset is larger than 5k rows, metrics shown in the Model performance tab of the explanation dashboard are computed on subsampled data, which may be different than the model metrics on the full dataset shown in the Metrics tab of the run.

Explainer: mimic.lightgbm

View previous dashboard experience

Model performance | Dataset explorer | Aggregate feature importance | Individual feature importance

DATA STATISTICS
Regressor
2220 datapoints
10 features

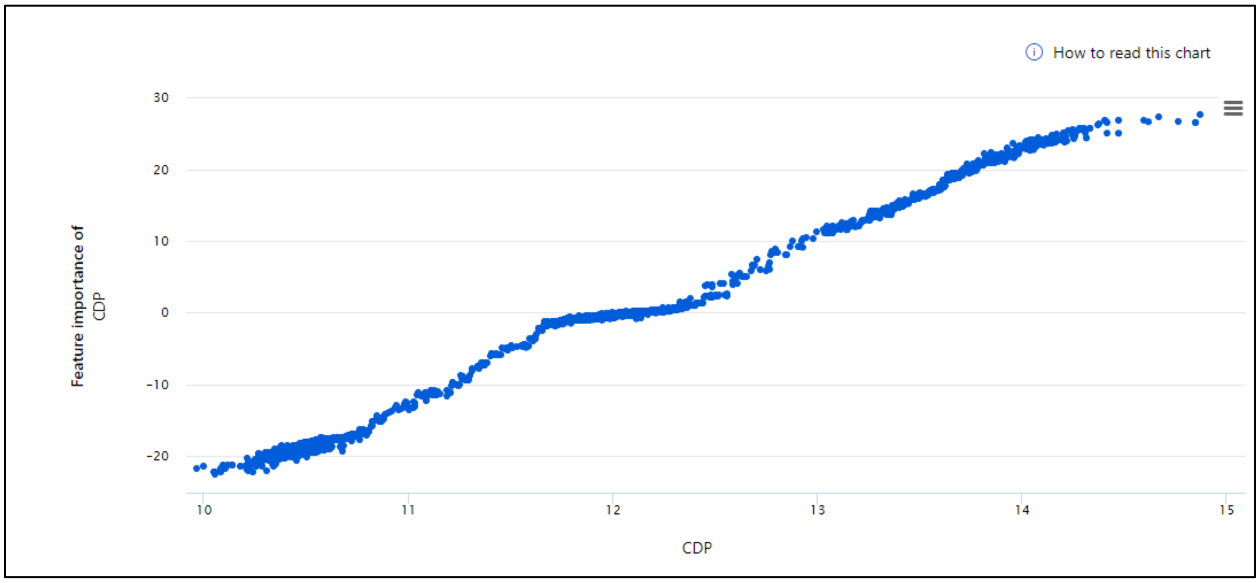
DATASET COHORTS
All data
2220 datapoints
0 filters

Explore the top-k important features that impact your overall model predictions (a.k.a. global explanation). Use the slider to show descending feature importances. All cohorts' feature importances are shown side by side and can be toggled off by selecting the cohort in the legend. Click on any of the features in the graph to see a density plot below of how values of the selected feature affect prediction.

Top 4 features by their importance

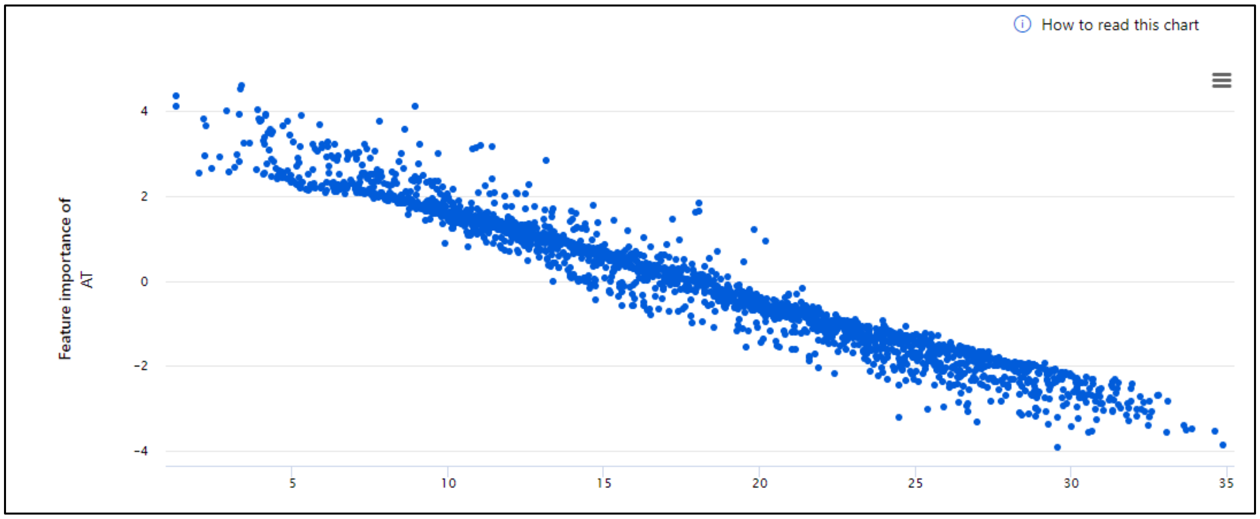
Sort by datapoint: All data
Chart type: [✓ Bar] Box

We can determine the 4 most important features on determining the Turbine Energy Yield, which are: Compressor discharge pressure (CDP), Turbine inlet temperature (TIT), Ambient temperature (AT), and Gas turbine exhaust pressure (GTEP). We can see that CO and NOx emissions aren't key features on calculating the Turbine Energy Yield, thus, the next steps would be on reducing the emissions without indirectly affecting the energy yield.



We can see the feature importance of the compressor discharge pressure; a higher discharge pressure will have more significance on determining the turbine energy yield. We see that this has the largest significance on the energy yield and should be prioritized when creating/maintaining a gas turbine.

Alexander Cheng – 1001634298
alexand.cheng@mail.utoronto.ca



We can also see the feature importance of the ambient temperature; a lower ambient temperature will have more significance on determining the turbine energy yield.