**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.**

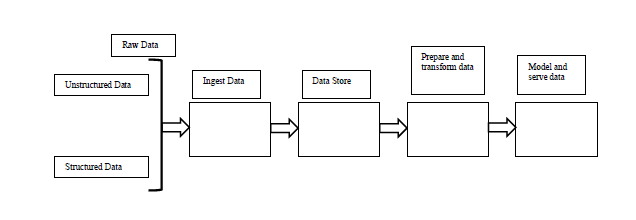
Azure Data Lake: Provides scalable storage for relational, semi-relational, and/or unstructured data

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.

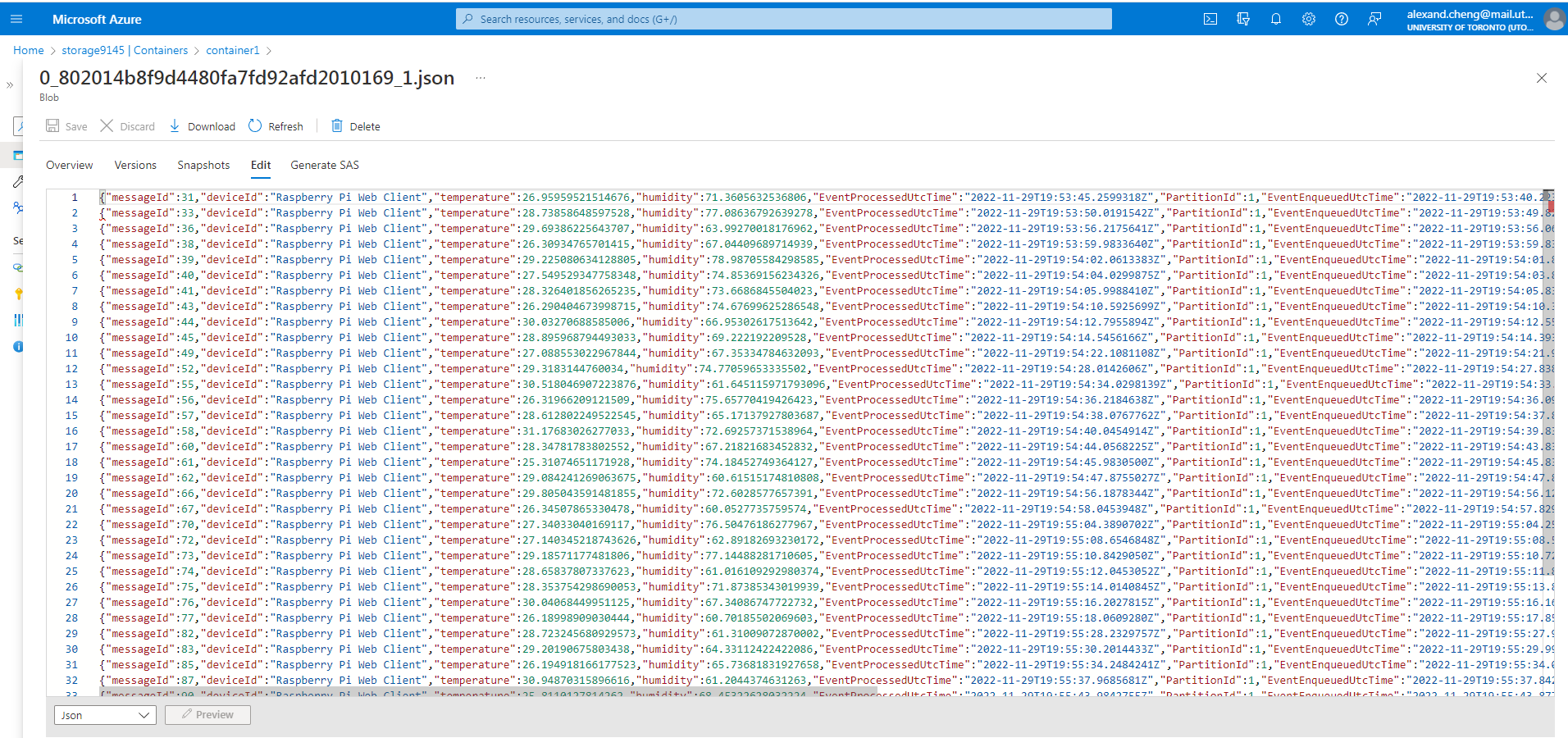
Prepare and Transform Data: **Azure Databricks** allows the preparation and the transformation (i.e. clean, sort, merge, join, etc.) of data in a notebook environment. It is used to prepare and transform the stored data.

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

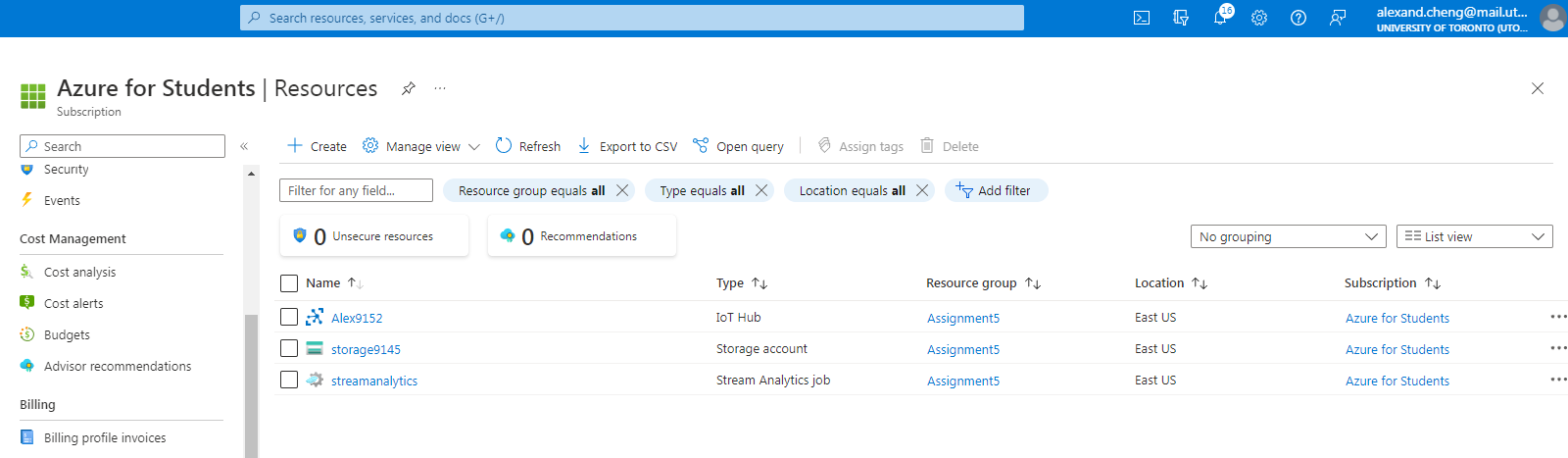
**Question 2: Explain how Stream Analytics works in Azure.**

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.

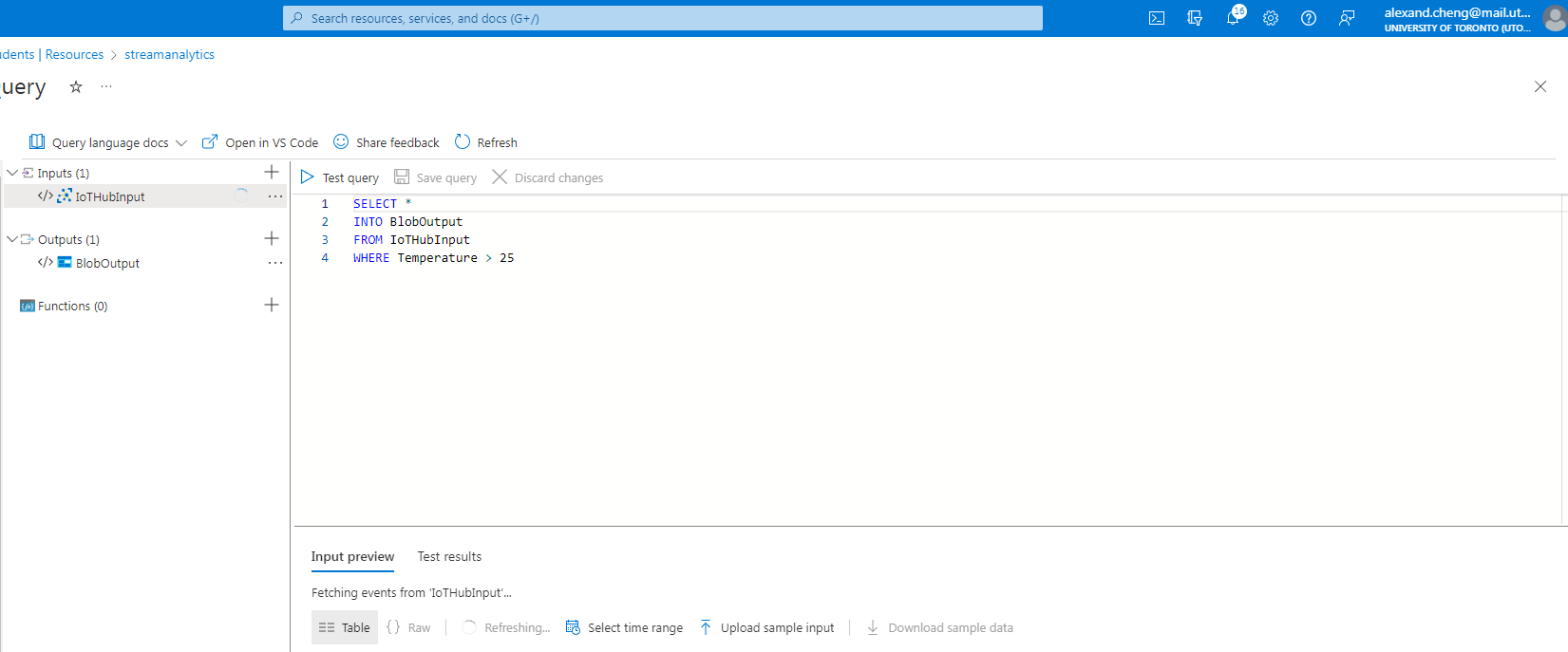
**Question 3: Implementing a Stream Analytics job by using the Azure portal**



Screenshot of the top 30 results



Deployed Resources



Screenshot of Query



Another screenshot of the top 30 results

**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.

1. Explain your dataset. Explore your dataset and provide at least 5 meaningful charts/graphs with an explanation.
2. 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

Graphical user interface, text, application, email

Description automatically generated

Deployed compute instance to run notebook on Azure Portal

Text

Description automatically generated with low confidence

Table

Description automatically generated with medium confidence

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

Table

Description automatically generated

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

**Correlation Plot**

Chart, treemap chart

Description automatically generated

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**

Chart

Description automatically generated

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

**Probability Plots**

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

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**

Chart, scatter chart

Description automatically generated

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.

**Gas Emissions and Global Warming Concerns**

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

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

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

**Diagram

Description automatically generated**

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.

**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.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

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

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

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.

1. 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.

Table

Description automatically generated

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.

Graphical user interface, text, application, email

Description automatically generated

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.

Graphical user interface, text, application

Description automatically generated

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.

Graphical user interface, text

Description automatically generated

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.

Graphical user interface, text, application, email

Description automatically generated

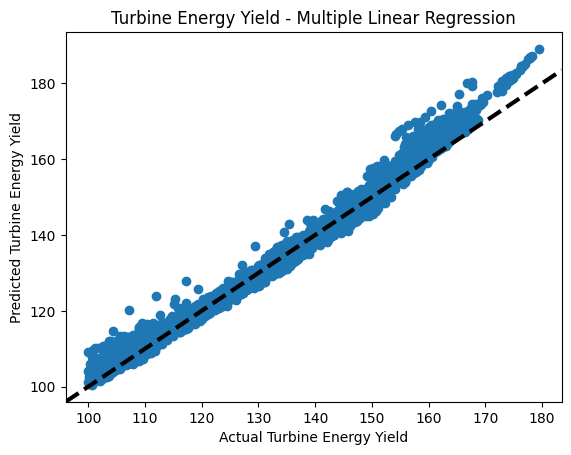
First tried this model with the learning rate as 0.1, and got the RMSE at around 46, this meant that the model was overfitting. Now the RMSE is around 4.395 MWH which is relatively low with a R2 score of 91.3% without too many hyperparameter tuning.

Graphical user interface, text, application, email

Description automatically generated

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**

****

**Chart, scatter chart

Description automatically generated**

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.

1. 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:**

Graphical user interface, text, application

Description automatically generated

AutoML Job for the Gas Turbine Dataset

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

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

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

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.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface

Description automatically generated

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

Graphical user interface

Description automatically generated with medium confidence

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.

Graphical user interface

Description automatically generated with medium confidence

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.

Graphical user interface, text, application

Description automatically generated

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.

Chart, line chart

Description automatically generated

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.

Chart, scatter chart

Description automatically generated

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.