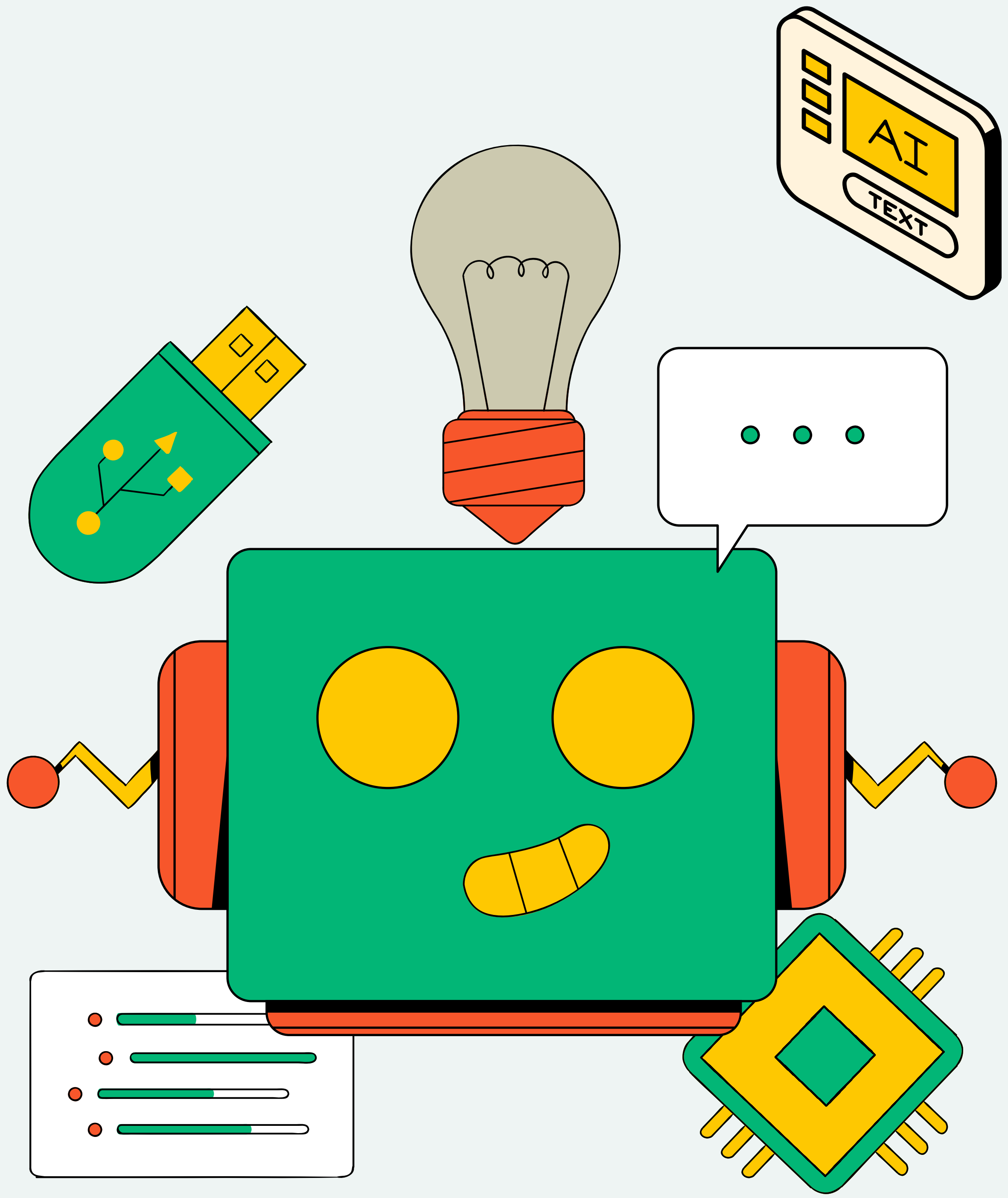
THYNK UNLIMITED

WE LEARN FOR THE FUTURE

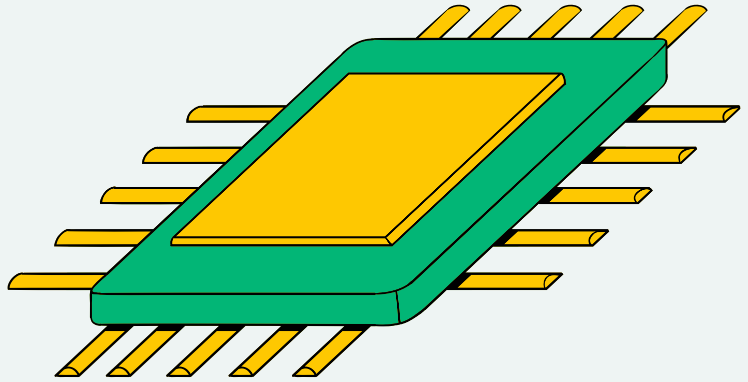


WIA1006 MACHINE LEARNING GROUP PROJECT

ELECTRICITY REVENUES PREDICTION

PRESENTED BY:

ABRACADABRA



TEAM MEMBERS

Saad Ahmed Pathan (22114077)

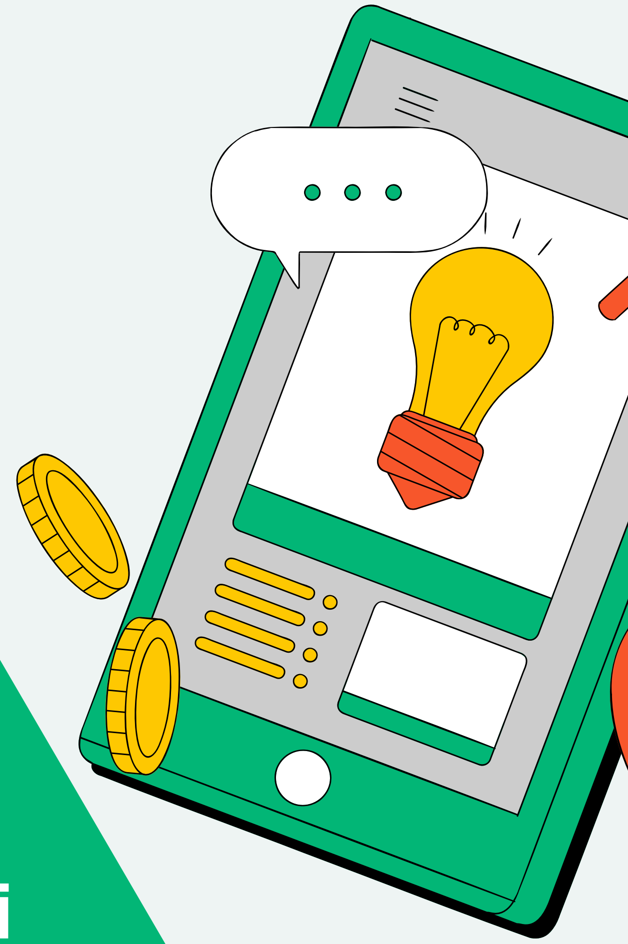
Samio Ayman (22082403)

Nurina Humaira Binti Mohd Romzan (22002204)

Nur Shaheila Ashriza Binti Mohd Saupi (22001745)

Nur Aina Batrisyia Binti Zakaria (23005013)

Siti Hajar Binti Mohd Nor Azman (22002035)



BACKGROUND OF THE PROBLEM & DATASET

- The dataset, titled "electricity_consumption_data.csv", contains detailed information about electricity utility, including **annual revenues, MWh sold and average number of customers across different categories.**
- The goal is to develop a machine learning model **capable of accurately forecasting electricity revenues based on the provided features.** This model is valuable for utility companies, energy firms, and policymakers who need to optimize electricity consumption.

- Number of samples: 1467
- Number of features: 15

- Source of data : The dataset is from the U.S. Energy Information Administration (EIA) on the website of Data.gov.

DATA PREPROCESSING

Identifying and handling missing values, outliers and inconsistencies

Encoding the categorical data

Feature scaling using PCA

EDA: SIGNIFICANT INSIGHTS

01

WHICH FEATURE WE THINK IS MOST IMPORTANT

Based on the exploratory data analysis (EDA), the feature **"Amount Sold for Residential in MWh" (ASforR)** appears to be the most important feature for predicting the target variable, Operating Revenues of Residential Sales (ORoRS). This is because:

- ASforR is directly related to ORoRS, as the amount of electricity sold to residential customers is a primary driver of revenue from residential sales.
- During the analysis, it is likely observed that ASforR has a high correlation coefficient with ORoRS, indicating its strong predictive power.

02

WHICH FEATURE TURNED OUT IS SUPRISINGLY IMPORTANT

A feature that turned out to be surprisingly important might be **"Average Number of Customers in Commercial & Industrial" (ANoCCI)**. It is not directly related to residential sales, but it may have surfaced as significant due to the following reasons:

- ANoCCI might influence the overall financial health and operational scale of the utility company, indirectly affecting residential sales operations.
- During EDA, it could be discovered that ANoCCI has a notable correlation with ORoRS, potentially due to shared operational efficiencies or customer base dynamics across different customer types.

EDA: HYPOTHESIS QUESTIONS

BEFORE ANALYSING THE DATASET (CONFIRMATORY DATA ANALYSIS)

- Correlation Hypothesis

Is there a significant correlation between the principal component (PC1) and the target variable (ORoRS)?

- Predictive Power Hypothesis

Can PC1, derived from PCA, significantly predict ORoRS using a linear regression model?

- Model Performance Hypothesis

Does a Random Forest Regressor outperform a linear regression model in predicting ORoRS?

- Overfitting Concern

Does tuning hyperparameters of the Random Forest Regressor or XGBoost model reduce overfitting and improve the generalizability of the model?

AFTER ANALYSING THE DATASET (EXPLORATORY DATA ANALYSIS)

- Unexpected Trends

What unexpected trends or patterns are present in the relationship between PC1 and ORoRS?

- Model Comparison Insights

How do different models compare in terms of performance metrics (MSE, RMSE, R-squared), and what does this tell us about the nature of the data?

- Data Distribution

How does the distribution of PC1 and ORoRS affect the model's predictions and performance?

- Dimensionality Reduction Impact

How does the use of PCA (particularly using only PC1) impact the model's ability to predict ORoRS compared to using the full set of features?

MODEL EVALUATION

MODEL	MSE	RMSE	R^2
LINEAR REGRESSION	231588760521001.2	15218040.626867875	0.9067015665821766
NEURAL NETWORK REGRESSION	222835706743011.4	14927682.564383911	0.9102278439510403
DECISION TREE	6411923390724.246	2532177.5985748405	0.997416876336296
RANDOM FOREST REGRESSOR	**5297160256238.563**	**2301556.051074699**	**0.9978659726302849**
XGBOOST REGRESSION	13617885819724.05	3690241.9730586843	0.9945138640986516

CONCLUSION

In conclusion, the **Random Forest Regressor** model is recommended for our project, as it exhibited the **best performance** in terms of predictive accuracy and model fit. It provided the lowest MSE and highest R-squared value among all models, indicating superior predictive capability.

However, depending on specific project requirements, the XGBoost Regression model could also be considered as it demonstrated strong performance as well.

The decision tree model, while showing promise, might require additional regularization techniques **to mitigate overfitting**.

The neural network and linear regression models **did not perform as well** and are less suitable for this dataset.