# DATA SCIENCE

# (CSE-558)

# Project: Energy Consumption Analytics

**GitHub Repository**
**Datasets Link**

## Group Members

Prem Kamal Jain — 2021483
Aryan Dhull — 2021520
Deepanshu — 2021524
Lakshya Kumar — 2021536
Prerak Gupta — 2021552

# 1. Datasets

## 1.1. Client Data

The Client Data contains a total of 146,070 data points with 26 features. Includes detailed information on customers, such as demographic details, electricity and gas consumption patterns, and contractual information. Demographic details, such as activation dates and contract end dates, provide insight into customer lifecycle stages. Consumption patterns include monthly energy usage, power capacity limits, and forecasted energy prices, helping to analyze customer behavior over time. In addition, the data set includes contractual details, such as the number of active products and contract types, which highlight customer engagement levels and service dependency.

The primary focus of this dataset is the target variable, *"Churn"*, which is binary in nature. A value of 1 indicates that the customer churned (left the service) within three months, while a value of 0 indicates that the customer was retained. This target variable serves as the foundation for predictive modeling to identify customers at risk of churn.

## 1.2. Price Data

The Price Data consists of 1,930,030 data points with 9 features. Provides context related to the external market by capturing time series energy prices across different periods, such as peak, mid-peak, and off-peak hours. These fluctuations in energy prices are critical to analyzing how pricing trends influence customer decisions. The data set also includes monthly energy price variations, which allow us to study seasonal trends and their impact on customer churn behavior.

Unlike the Client Data, the Price Data does not contain a target variable. Instead, it enriches the analysis by offering contextual information about the energy market, making it possible to measure price sensitivity and correlate price variability with churn rates.

## 1.3. Merging the Datasets

The process began with grouping and aggregating the price data to calculate average off-peak prices by `id` and `price_date`. Following this, price differences were calculated to capture seasonal variations. Specifically, records for January and December were extracted for each `id`, and the off-peak price difference between these two months was calculated and introduced as an additional feature. Finally, the calculated price differences were merged with the client dataset (*146,070 points with 64 features*) using the `id` column to enrich the dataset with valuable price sensitivity information, enhancing its relevance for downstream analysis.

# 2. Problem Statement

PowerCo, a major utility provider, is facing a significant challenge due to the increase in customer turnover in an increasingly competitive energy market. Customers have the flexibility to switch providers based on pricing, service quality, and competitive offers. This situation has led to revenue losses, operational inefficiencies, and difficulty in sustaining customer loyalty.

The primary objective of this project is to identify the factors that contribute to customer churn and provide actionable insights to improve customer retention. Specifically, the project aims to analyze customer demographics, consumption behaviors, and pricing data to determine which features have the most significant influence on churn. In addition, it seeks to quantify the role of price sensitivity in customer decisions by studying the effect of energy price fluctuations during peak and off-peak periods.

By achieving these objectives, the project aims to develop a predictive framework that can identify customers at risk of churn and propose strategies to mitigate it. These insights will help PowerCo minimize churn, improve customer satisfaction, and optimize resource allocation to focus on retention initiatives.

# 3. Challenges Faced and Solutions

While working with the dataset, several challenges were encountered that required careful handling to ensure high-quality data for analysis and modeling. The key challenges and corresponding solutions are described in the following.

## 3.1. Inconsistencies in the Data

*Challenge:* The raw dataset contained errors and inconsistencies, such as invalid values, duplicate records, or formatting problems. These inconsistencies could have led to inaccurate results or unreliable model performance.

*Solution:* To address these issues:

- We carefully identified and resolved anomalies in the data.
- Errors were corrected and inconsistencies were removed to ensure data accuracy.

By cleaning the data, we reduced noise and improved the overall reliability of the dataset.

## 3.2. Missing Values

*Challenge:* Missing values were present in critical fields of the dataset, which could introduce bias during model training or lead to incomplete analyses.

*Solution:* To handle missing values effectively:

- *Imputation techniques* were applied to fill missing entries where possible (e.g., using averages, medians, or other relevant methods).

- In cases where imputation was not feasible, incomplete records were removed to maintain data integrity.

These steps ensured the dataset was complete and minimized the risk of bias in the models.

### 3.3. Feature Engineering

*Challenge:* The raw dataset lacked sufficient features to capture underlying patterns and insights, limiting the model's ability to learn effectively. Additionally, some features were not in the optimal format for modeling.

*Solution:* To address these limitations:

- New features were *created* from the existing data to capture more insights and improve predictive power.
- Data was *transformed* into suitable formats for modeling, ensuring compatibility with machine learning algorithms.

Feature engineering enhanced the quality and richness of the dataset, enabling better performance and interpretability of the models.

### 3.4. Conclusion

By overcoming these challenges, we ensured that the dataset was clean, complete, and suitable for building effective and reliable models. Each step significantly contributed to improving the accuracy, consistency, and overall quality of the analysis.

## 4. Hypothesis Test and their Insights

### 4.1. Discount and Churn

**Hypothesis:**

- **Null Hypothesis ($\mathbf{H}_0$):** Discounts have no significant effect on churn rates.
- **Alternative Hypothesis ($\mathbf{H}_1$):** Discounts significantly affect churn rates.

**Methodology:** Conducted a *t*-test to analyze the correlation between forecast discounts and churn.

**Results:**

- Correlation Coefficient ($r$): $0.017$ (negligible correlation).
- T-Statistic: $2.058$ (below critical value $2.576$ at $\alpha = 0.01$).

**Conclusion:** Discounts had minimal impact on churn rates, suggesting that monetary incentives alone may not be sufficient to retain customers.

### 4.2. Price Variability and Churn

**Hypothesis:**

- **Null Hypothesis ($\mathbf{H}_0$):** Price variability does not impact churn.
- **Alternative Hypothesis ($\mathbf{H}_1$):** Price variability significantly affects churn.

**Results:**

- T-Statistic: $-2.889$
- P-Value: $0.00387$ (significant at $\alpha = 0.05$).

**Conclusion:** Customers who churned exhibited higher sensitivity to price variability, highlighting the need to stabilize pricing strategies.

### 4.3. Tenure and Churn

**Hypothesis:**

- **Null Hypothesis ($\mathbf{H}_0$):** Tenure does not influence churn.
- **Alternative Hypothesis ($\mathbf{H}_1$):** Tenure significantly affects churn.

**Results:**

- T-Statistic: $-8.77$
- P-Value: $1.92 \times 10^{-18}$ (highly significant).

**Conclusion:** Shorter-tenured customers were more likely to churn, emphasizing the importance of early engagement strategies.

### 4.4. Seasonal Price and Churn

**Hypothesis:**

- **Null Hypothesis ($\mathbf{H}_0$):** There is no significant difference in seasonal price changes between churned and non-churned customers.
- **Alternative Hypothesis ($\mathbf{H}_1$):** There is a significant difference in seasonal price changes between churned and non-churned customers.

**Methodology:** Conducted an independent t-test to compare the mean values of *off_peak_peak_var_mean_diff* (a measure of seasonal price changes) between churned and non-churned customers.

**Results:**

- T-Statistic: $-3.217$
- P-Value: $0.0013$ (for significance level $\alpha = 0.05$).

**Conclusion:** There is a significant difference in seasonal price changes between churned and non-churned customers. Seasonal price fluctuations likely play a role in customer churn.

### 4.5. Customer Support Frequency and Churn

**Hypothesis:**

- **Null Hypothesis ($H_0$):** There is no significant difference in the average customer support interaction frequency between churned and non-churned customers.
- **Alternative Hypothesis ($H_1$):** There is a significant difference in customer support interaction frequency between churned and non-churned customers, implying that higher interaction frequency may impact retention.

**Methodology:** Conducted an independent t-test to compare the mean values of customer support interaction frequency between churned and non-churned customers.

**Results:**

- T-Statistic: $-3.01$
- P-Value: $0.0026$ (for significance level $\alpha = 0.05$).

**Conclusion:** There is a significant difference in customer support interaction frequency between churned and non-churned customers, suggesting that higher frequency may be linked to churn.

### 4.6. Product Variety and Churn

**Hypothesis:** Clients with a broader array of products or services might be less likely to churn due to increased engagement.

**Null Hypothesis ($H_0$):** There is no significant difference in churn rates between customers with a single product and those with multiple products.

**Alternative Hypothesis ($H_1$):** Customers with multiple products have a significantly lower churn rate than those with only one product.

**Methodology:** The methodology here involves using a Chi-square test to assess the relationship between product variety (single vs. multiple products) and churn rates.

**Results:**

- Chi2-Statistic: $4.11754407914068$
- Critical Value: $6.635$ (for significance level $\alpha = 0.01$)

**Conclusion:** There is no significant difference in churn rates between customers with a single product and those with multiple products.

## 5. Modelling: Random Forest Classifier

### 5.1. Why Random Forest Was Chosen

- Handling Non-Linear Relationships: Random Forest can model complex interactions between features, making it suitable for analyzing customer churn driven by multiple factors like pricing, tenure, and consumption.
- Robustness to Overfitting: By combining multiple decision trees, Random Forest minimizes overfitting, ensuring better generalization on unseen data.
- Feature Importance Analysis: Random Forest ranks feature importance, providing insights into the drivers of churn, such as price sensitivity and tenure.
- No Need for Feature Scaling: Unlike models like logistic regression or SVM, Random Forest operates effectively on unscaled data, saving preprocessing time.

### 5.2. Specifications of the Model

- Algorithm: Random Forest Classifier
- Training/Test Split: 75
- Hyperparameters:
  - Number of Trees (n_estimators): 1000
  - Maximum Depth: None (trees grow until all leaves are pure).
  - Minimum Samples Split: 2
  - Minimum Samples per Leaf: 1

### 5.3. Results and Evaluation

- Accuracy: 89.87%
- Precision: 45.23%
- Recall: 5.19%
- Training Time: 341.65 seconds

### 5.4. Key Observations

- **Accuracy (89.87%)**: Indicates that the model correctly predicted most outcomes, but high accuracy can be misleading in imbalanced datasets.
- **Precision (45.23%)**: Reflects the model's ability to correctly identify churners among predicted churn cases.
- **Recall (5.19%)**: Highlights the model's struggle to identify actual churners, missing a significant number of them.
- **Training Time**: Computationally intensive due to the high number of estimators.

### 5.5. Implications of Results

- **Low Recall**: The model underperformed in detecting churn cases, suggesting the need for additional techniques like oversampling, boosting algorithms (e.g., XGBoost), or incorporating more informative features.
- **Feature Insights**: Features like net margin, price variability, and tenure significantly influenced churn predictions, guiding business decisions.

## 6. Scaling Techniques

### 6.1. Techniques Used

- **Principal Component Analysis (PCA)**:
  - Concept: Reduces dimensionality by projecting data onto principal components that explain the most variance.
  - Benefits: Reduces overfitting and computational cost.
  - Results:
    - Training Time: 201.3 seconds
    - Accuracy: 89
    - Precision: 38
    - Recall: 4
- **Singular Value Decomposition (SVD)**:
  - Concept: Factorizes the dataset into singular vectors, isolating dominant components.
  - Benefits: Maintains variance while reducing complexity.
  - Results:
    - Training Time: 222.16 seconds
    - Accuracy: 90
    - Precision: 54
    - Recall: 10
- **Random Projection**:
  - Concept: Maps high-dimensional data into lower-dimensional space using random matrices.
  - Benefits: Computationally efficient.
  - Results:
    - Training Time: 220.63 seconds
    - Accuracy: 89
    - Precision: 45
    - Recall: 5
- **JL-Lemma (Johnson-Lindenstrauss Lemma)**:
  - Concept: Preserves pairwise distances during dimensionality reduction, maintaining data structure.
  - Benefits: Effective in preserving geometrical relationships.
  - Results:
    - Training Time: 618.94 seconds
    - Accuracy: 90
    - Precision: 69
    - Recall: 8

### 6.2. Analysis of Scaling Techniques

- **Best Balance (SVD)**: Maintained high accuracy (90%) while reducing training time significantly.
- **Fastest (PCA)**: Achieved the shortest training time but sacrificed recall (4%).
- **Most Accurate (JL-Lemma)**: Delivered the highest precision (69%) but at a high computational cost.

### 6.3. Conclusion on Scaling Techniques

- Scaling techniques, particularly SVD, effectively reduced computational costs without compromising model performance, making them valuable in handling high-dimensional datasets.

## 7. Conclusion and Future Work

### 7.1. Conclusion

The Random Forest model achieved 89.87% accuracy and excelled at identifying non-churners, which is useful for understanding customer base stability. However, its low recall (5.19%) indicated poor performance in detecting actual churners, leading to a significant number of false negatives. Statistical tests revealed that factors such as price variability, tenure, and seasonal price changes were the most significant drivers influencing churn. These findings highlight the complexity of churn prediction and the challenges in balancing model performance.

### 7.2. Future Directions

- **Enhanced Feature Engineering**: Incorporate additional behavioral data such as payment histories and customer complaints to gain deeper insights into customer behaviors and churn patterns.
- **Addressing Class Imbalance**: Explore advanced resampling techniques like ADASYN and ensemble undersampling to balance the dataset and improve recall.
- **Algorithmic Exploration**: Experiment with gradient boosting techniques such as XGBoost, LightGBM, and CatBoost to potentially increase recall and model performance.
- **Time-Series Analysis**: Use temporal features (e.g., seasonal changes, monthly trends) to better predict churn trends over different periods.
- **Customer Segmentation**: Develop churn prediction models tailored to specific customer segments to improve precision and relevance of predictions.

By addressing these areas, the model can better support PowerCo's customer retention strategy and provide actionable business insights, ultimately driving customer satisfaction and reducing churn rates effectively.