Predicting Energy Consumption for Buildings at the University of Toronto

Report by Zach Liu on 06/12/2024

Detailed Process can be found: Colab File

1. Introduction and Motivation

This project, provided by the Sustainability Office at the University of Toronto, aims to predict whether a building's monthly energy consumption will exceed its baseline. The goal is to assist in the efficient management of energy consumption, which is crucial for both economic and environmental sustainability. The decision to use a regression-based model, specifically logistic regression, was motivated by the need to model a binary outcome (energy consumption exceeding baseline or not). Logistic regression was chosen due to its simplicity, interpretability, and effectiveness in binary classification problems.

2. Data Collection and Preprocessing

Data Sources:

- **Weather Data:** https://toronto.weatherstats.ca/download.html
- **Energy Data:** Provided by the Sustainability office (04/2018 - 03/2021)

Steps Taken:

1. **Loading Data:**
   - Weather data was loaded from a CSV file.
   - Energy data was loaded from an Excel file.
2. **Date Conversion:**
   - Converted date columns to datetime format for both datasets.
3. **Filtering and Matching Date Ranges:**
   - Filtered the weather data to match the date range of the energy data, ensuring consistency.
4. **Aggregation:**
   - Aggregated daily weather data to monthly averages to match the granularity of the energy data.
5. **Merging Data:**
   - Merged weather and energy data on the month field.

3. Feature Engineering

Key Features:

- **Weather-Related Features:**
  - max_temperature_v
  - min_temperature_v
  - max_relative_humidity_v
  - min_relative_humidity_v
  - precipitation_v
  - rain_v
  - snow_v
- **Temporal Feature:**
  - month_number (month of the year)

Additional Steps:

- **Baseline Calculation:**
  - Calculated the baseline energy consumption for each building as the mean consumption over the entire period.
- **Target Variable:**
  - Created a binary target variable indicating whether the consumption in a given month exceeds the building's baseline consumption.

4. Handling Class Imbalance

Technique Used:

- **SMOTE (Synthetic Minority Over-sampling Technique):**
  - Applied SMOTE to oversample the minority class (months where consumption exceeds baseline).

5. Model Training and Evaluation

Models Used:

1. **Logistic Regression:**
   - Initially used to establish a baseline performance.
   - Increased max_iter and used the 'saga' solver to handle convergence issues.
2. **Random Forest:**
   - Trained with hyperparameter tuning using Grid Search to find optimal parameters.

3. **Gradient Boosting:**
    - Included as part of the ensemble for its robustness and ability to handle non-linear relationships.
4. **Ensemble Model:**
    - Combined Logistic Regression, Random Forest, and Gradient Boosting in a Voting Classifier to leverage the strengths of each model.

Hyperparameter Tuning:

- Performed Grid Search to optimize parameters for Random Forest and Gradient Boosting models.

6. Results

Model Performance:

- **Logistic Regression:**
    - Achieved balanced performance with precision and recall around 0.67 after addressing convergence issues.
- **Random Forest:**
    - Best parameters: {'max_depth': None, 'min_samples_split': 2, 'n_estimators': 300}
    - Precision: 0.67, Recall: 0.66, Accuracy: 67%
- **Ensemble Model:**
    - Achieved 67% accuracy with balanced precision and recall for both classes.

Feature Importance:

- Top features influencing the predictions were:
    - max_temperature_v
    - min_temperature_v
    - max_relative_humidity_v
    - min_relative_humidity_v
    - precipitation_v
    - rain_v
    - snow_v
    - month_number

7. Conclusion

This project successfully demonstrated the application of data science and machine learning techniques to predict energy consumption for university buildings. By incorporating weather data and leveraging advanced models, we achieved a balanced and robust prediction model. However, the accuracy of the model can be improved by training with more data. The currently model is not recommended for employment. The insights gained from feature importance analysis can help facility managers focus on key factors influencing energy usage, enabling more effective energy management strategies.

8. Future Work

- **Advanced Feature Engineering:**
  - Incorporate more sophisticated features like interaction terms, polynomial features, and lagged variables.
- **Further Model Optimization:**
  - Explore other ensemble techniques like stacking and boosting.
- **Real-Time Implementation:**
  - Develop a real-time prediction system to provide ongoing insights and alerts for energy management.
- **Cross-Validation:**
  - Implement cross-validation to ensure the robustness and generalizability of the model.

9. Acknowledgments

Thanks to the Sustainability Office at the University of Toronto for providing the data and the resources necessary to conduct this project. Special thanks to the data science community for the valuable tools and libraries that made this analysis possible.