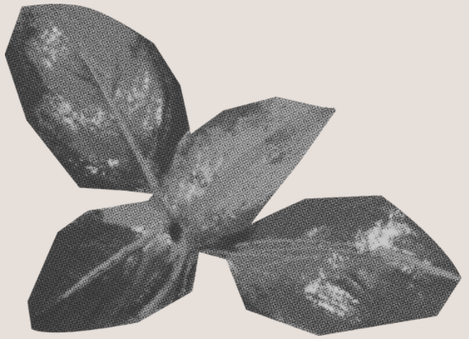
DATA 607 PROJECT PRESENTATION

SUSTAINABILITY ANALYSIS OF COMMERCIAL BUILDING ENERGY CONSUMPTION IN THE UNITED STATES

Presented by: Annie Shamirian, Osama Bilgrami & Umair Qureshi

INTRODUCTION



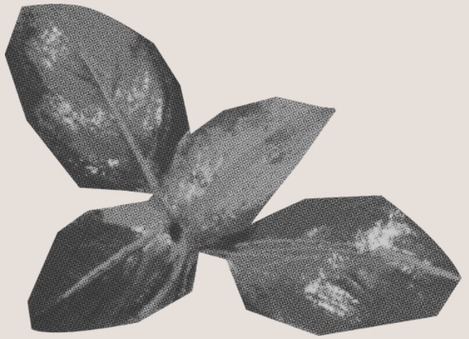
1. A study of commercial building energy consumption

2. The dataset used is the 2018 CBECS Survey Data from the U.S. Energy Information Administration

3. Contains data on building features and energy usage from commercial buildings across 50 U.S. states and the District of Columbia

4. Sample size of 6436 with 135 features

OBJECTIVES



We aim to predict energy consumption by analyzing building characteristics to understand sustainability factors and answer the following key questions:

1. Which features are the most important in determining total energy consumption?

2. What is the magnitude of influence each important feature has on total energy consumption?

DATA PREPARATION

- Dropped derived and modified variables
- Checked for outliers and dropped nulls, imputation flag columns and weight columns because they don't add to predictive power
- Checked for VIF (Variance Inflation Factor) and dropped high VIF features
- Created a new target variable from total energy consumption using median to build a classification model
- Greater than median classified as 1, and 0 if lower

PREDICTIVE MODELING

7 MODELS:

SVM
QDA
NAIVE BAYES
LOGISTIC REGRESSION
DECISION TREE
RANDOM FOREST
LIGHTGBM

Our pipeline involved the following steps:

- Splitting data into 80% training and 20% testing data
- Data transformation using PowerTransFormer
- Feature scaling with StandardScaler
- HyperParameter Tuning using GridSearchCV and 5-fold-cross-validation
- Choosing the model with the highest accuracy
- Employing best model to predict target values for the test set
- Model evaluation using classification report with accuracy, precision, recall, and F1- score metrics and confusion matrix

NAIVE BAYES

Naive Bayes model achieved classification accuracy of approximately 82.5%.

	precision	recall	f1-score	support
0	0.84	0.82	0.83	660
1	0.81	0.83	0.82	612
accuracy			0.83	1272
macro avg	0.83	0.83	0.83	1272
weighted avg	0.83	0.83	0.83	1272

QUADRATIC DISCRIMINANT ANALYSIS

QDA model achieved classification accuracy of approximately 85.1%

	precision	recall	f1-score	support
0	0.88	0.83	0.85	660
1	0.83	0.87	0.85	612
accuracy			0.85	1272
macro avg	0.85	0.85	0.85	1272
weighted avg	0.85	0.85	0.85	1272

SUPPORT VECTOR MACHINE

SVM model achieved classification accuracy of approximately 91.03%

	precision	recall	f1-score	support
0	0.92	0.91	0.91	660
1	0.90	0.91	0.91	612
accuracy			0.91	1272
macro avg	0.91	0.91	0.91	1272
weighted avg	0.91	0.91	0.91	1272

LOGISTIC REGRESSION

Logistic Regression model achieved an accuracy of approximately 91.67%

	precision	recall	f1-score	support
0	0.93	0.91	0.92	660
1	0.90	0.92	0.91	612
accuracy			0.92	1272
macro avg	0.92	0.92	0.92	1272
weighted avg	0.92	0.92	0.92	1272

RANDOM FOREST

Random Forest model achieved a classification accuracy of approximately 90.7%

	precision	recall	f1-score	support
0	0.92	0.89	0.91	660
1	0.89	0.92	0.91	612
accuracy			0.91	1272
macro avg	0.91	0.91	0.91	1272
eighted avg	0.91	0.91	0.91	1272

DECISION TREE

The DT model achieved a classification accuracy of approximately 89.85%

	precision	recall	f1-score	support
0	0.90	0.90	0.90	660
1	0.89	0.90	0.89	612
accuracy			0.90	1272
macro avg	0.90	0.90	0.90	1272
weighted avg	0.90	0.90	0.90	1272

LIGHT GBM

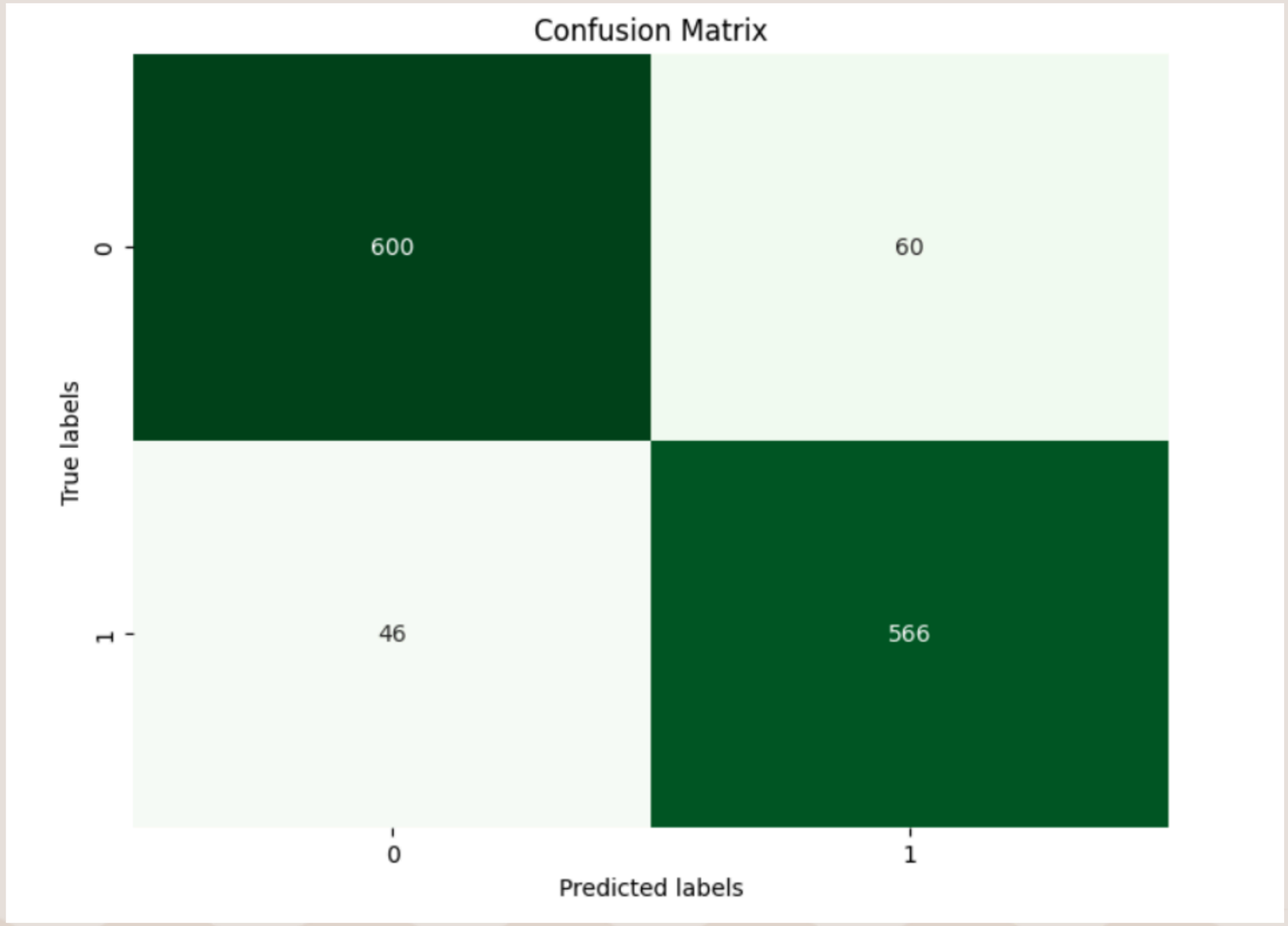
The LGBM model achieved a classification accuracy of approximately 91.67%

	precision	recall	f1-score	support
0	0.92	0.92	0.92	660
1	0.91	0.92	0.91	612
accuracy			0.92	1272
macro avg	0.92	0.92	0.92	1272
weighted avg	0.92	0.92	0.92	1272

BEST MODEL

We decided to choose the Logistic Regression model due to the following reasons:

- Comparable Accuracy to other models
- Superior Interpretability
- Transparent Decision-Making
- Effective Communication



IMPORTANT FEATURES

- Higher SQFT and more workers increase energy use.
- Warmer regions consume less energy, likely due to heating costs outweighing cooling expenses.
- Number of heating days impacts energy usage significantly.
- Narrow rectangular buildings are less energy-efficient.
- Vacant buildings and food service buildings lead to higher energy consumption.

QUESTIONS?

THANK YOU!