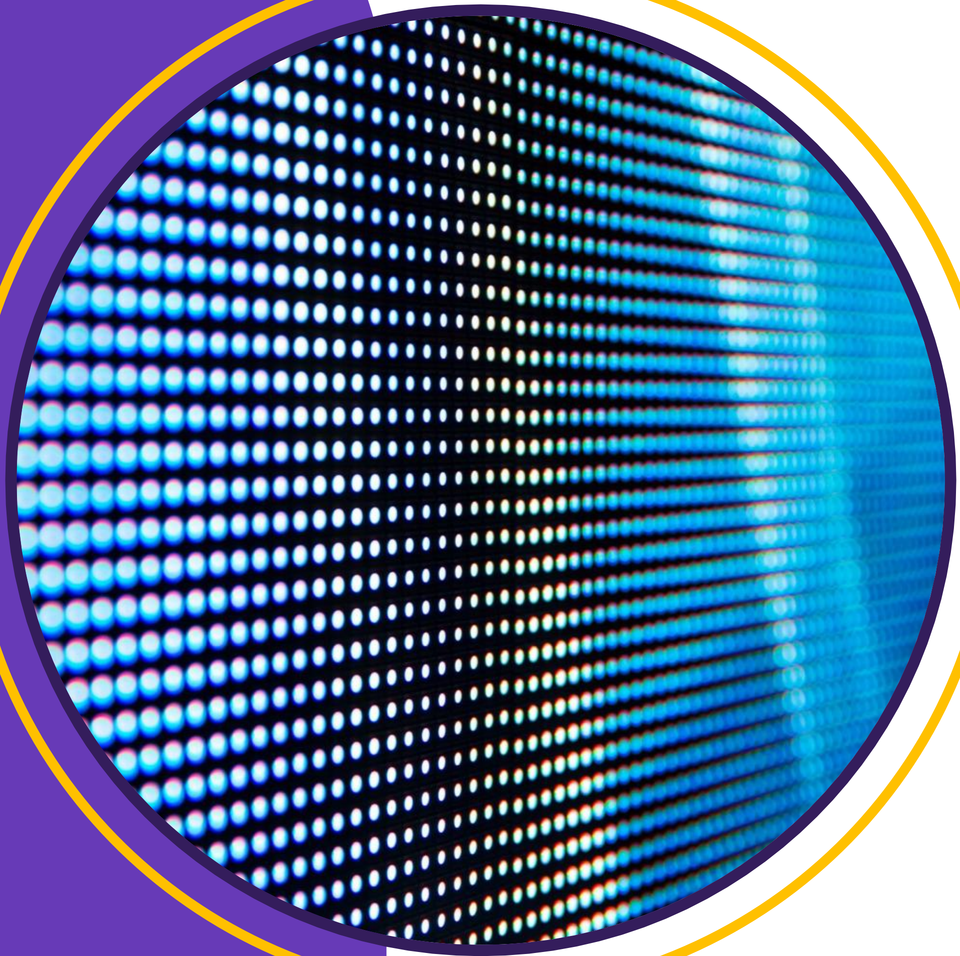
BUILDING ENERGY SAVINGS GUIDE:

Spotting Best Heating Upgrade Opportunities

NAME:	ELVIS KIPRONO
COURSE:	DATA SCIENCE – FULL TIME REMOTE
INSTITUTION:	MORINGA SCHOOLS
PHASE:	PHASE 3



Introduction

Project Overview:
This project leverages the U.S. Energy Information Administration's (EIA) Commercial Buildings Energy Consumption Survey (CBECS) 2018 dataset. Our primary goal is to identify commercial buildings that exhibit significant, targeted energy efficiency retrofit potential specifically for heating systems.

Problem Statement:
To pinpoint where heating energy efficiency improvements are most critically needed, we use the CBECS dataset to identify commercial buildings that consume unusually high amounts of energy in their heating systems when compared to similar buildings.

Motivation

- Identifying buildings ripe for heating efficiency upgrades is crucial for reducing overall energy use, operating expenses, and environmental impact.
- Traditional methods often focus on broad energy consumption intensity. However, a building might have an average overall energy use but be highly inefficient in a specific area like heating.
- By targeting these specific inefficiencies, this project aims to provide actionable insights for more focused and successful retrofit efforts.

Dataset Overview & Key Variables

Target Variable

01 HIGH_HEATING_POTENTIAL

02 HIGH_COOLING_POTENTIAL

03 HIGH_LIGHTING_POTENTIAL

DATASET

Commercial Buildings Energy Consumption Survey (CBECS) 2018 from EIA

Feature Variables

01 Binned Square Footage (SQFT_BINNED)

02 Heating Degree Days (HDD65)

03 Roof Construction (RFCNS)

04 Weekly Operating Hours (WKHRS)

05 Wall Construction (WLCNS)

06 Cooling Degree Days (CDD65)

07 Occupancy Type (OCCUPYP)

08 Year Constructed (YRCONC)

09 Number of Floors (NFLOOR)

MODEL: DECISION TREES & METHODOLOGY

- 1. Data Acquisition & Loading: Loaded CBECS 2018 public-use microdata.
- 2. Exploratory Data Analysis (EDA): Examined dataset structure, variables, distributions.
- 3. Feature Engineering: Calculated total heating energy, defined peer groups, created HIGH_HEATING_POTENTIAL.
- 4. Data Preprocessing: Handled missing values, converted categorical to numerical (one-hot encoding), scaled numerics, selected impactful features.
- 5. Model Training & Evaluation: Split data, applied SMOTE for imbalance, trained Decision Tree Classifier, evaluated with reports, ROC curves, confusion matrices.
- 6. Interpretation: Visualized Decision Tree for understanding its logic.

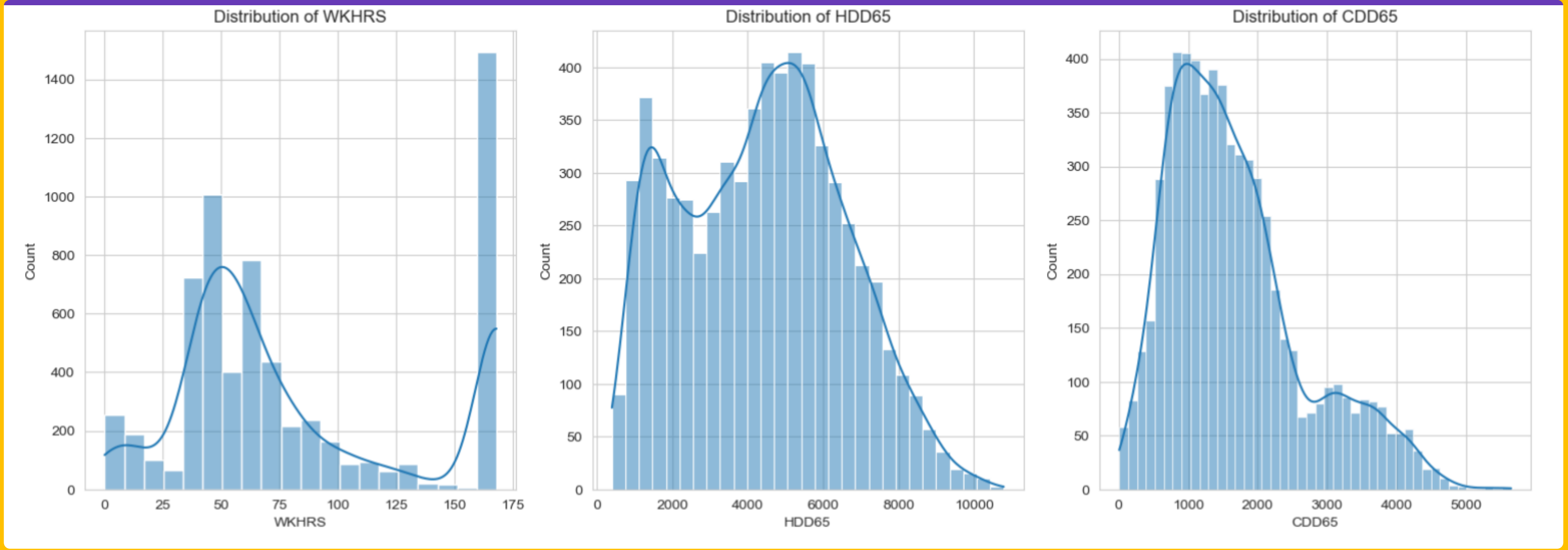
EDA: Key Numerical Features Distributions

Understanding these distributions confirms the characteristics of our input data and their potential influence on heating consumption.

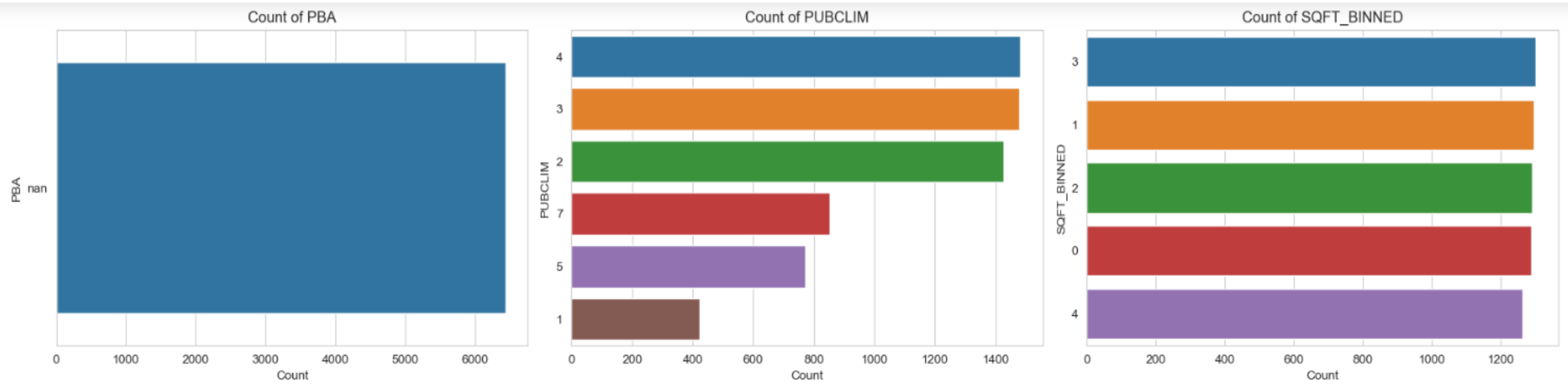
1

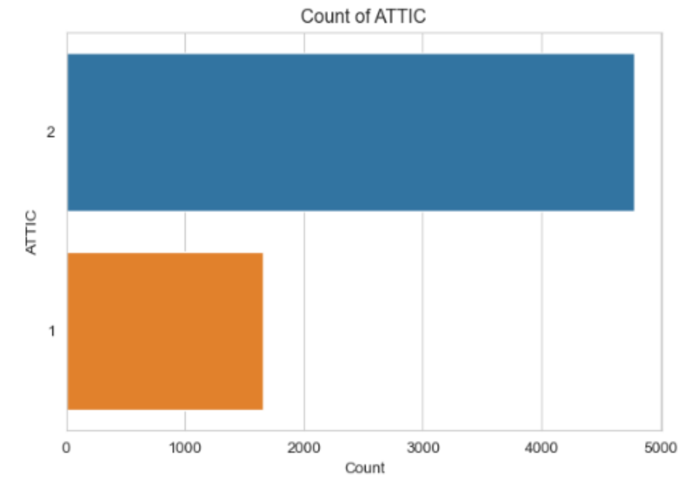
2

3

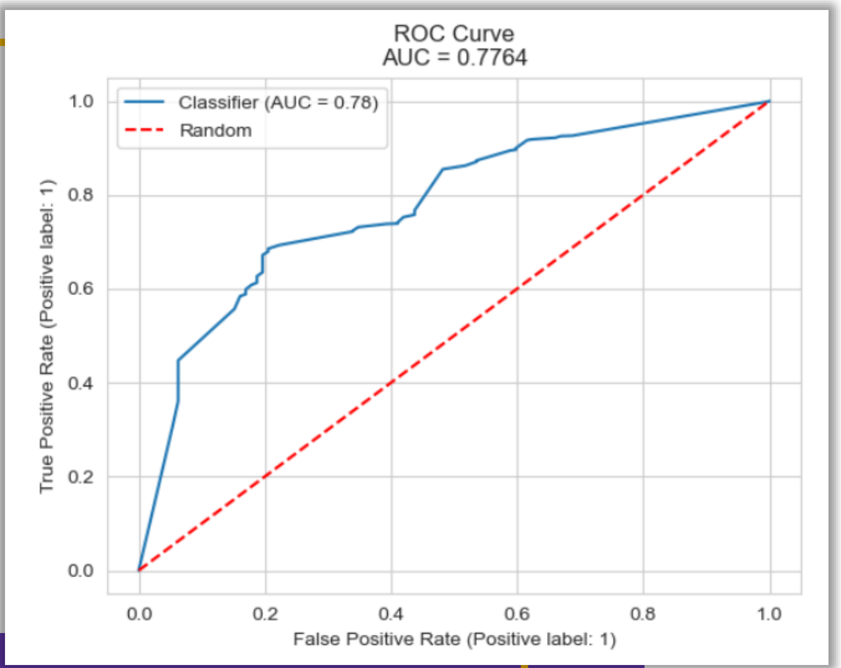


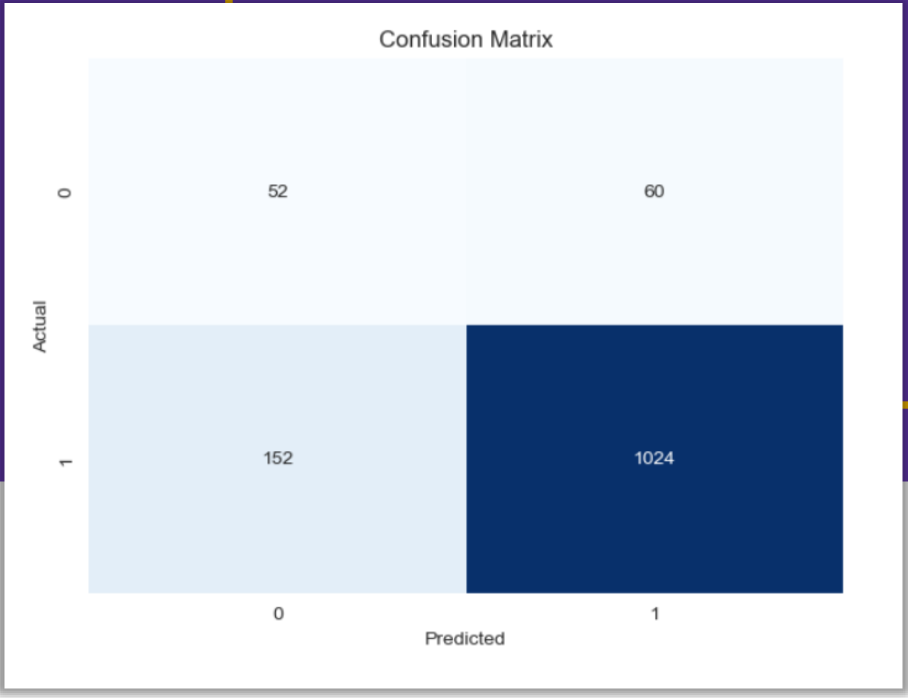
EDA: Key Categorical Features Distributions





Count plots illustrate the composition of categorical features such as Principal Building Activity (PBA), Public Climate Region (PUBCLIM), Binned Square Footage (SQFT_BINNED), and Attic presence (ATTIC). These distributions are essential for understanding the dataset's composition and for guiding one-hot encoding during preprocessing.





Model Performance:
Decision Tree Classifier

ROC Curve (AUC = 0.7760):
- Indicates a fairly good ability to distinguish between high and low heating potential buildings (better than random).

Confusion Matrix Interpretation:
- True Positives: 1022 buildings correctly identified as having high heating potential.
- False Negatives: 154 buildings with actual high heating potential that were missed (key area for future improvement).
- True Negatives: 52 buildings correctly identified as NOT having high heating potential.
- False Positives: 60 buildings incorrectly flagged as having high heating potential (false alarms).

RECOMMENDATION

1. **Prioritized Outreach**: Use model predictions to target high-potential buildings for retrofits first, maximizing ROI.
2. **Tailored Solutions**: Follow up with targeted energy audits based on model insights (e.g., roof/wall construction) to propose precise solutions.
3. **Data-Driven Decisions**: Leverage the model for informed investment decisions and to demonstrate tangible energy savings.
4. **Addressing Missed Opportunities**: Explore strategies to reduce false negatives in future model iterations to capture more high-value targets.
5. **Long-Term Strategy**: Integrate this model as a foundational element for continuous energy management and carbon footprint reduction.

CONCLUSION

Through a comprehensive process of data wrangling, feature engineering, and model training with a Decision Tree Classifier, we've demonstrated the ability to pinpoint buildings that stand to gain the most from energy efficiency upgrades. This data-driven approach provides valuable insights for stakeholders to prioritize interventions, optimize resource allocation, and contribute to significant reductions in energy consumption and operational costs.

Thank You!