# NYC Buildings Energy Consumption Prediction, Year 2021

Capstone Project - GA DSI 123

# CONTENTS



# Background


01 Grid Planning & Operation

02 Energy Policy & Regulation

03 Resource Allocation & Investment

04 Building Energy Management

05 Climate Resilience Planning

# Problem Statement

Develop a machine learning model that can **predict annual energy consumption** of New York City buildings

# Data Exploration

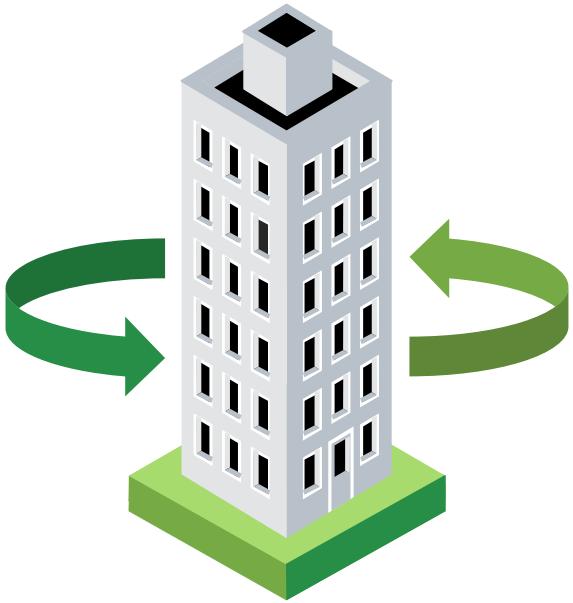
29K+ NYC buildings with 25K+ SQFt, 254 features, available since 2013

**Building Specs**

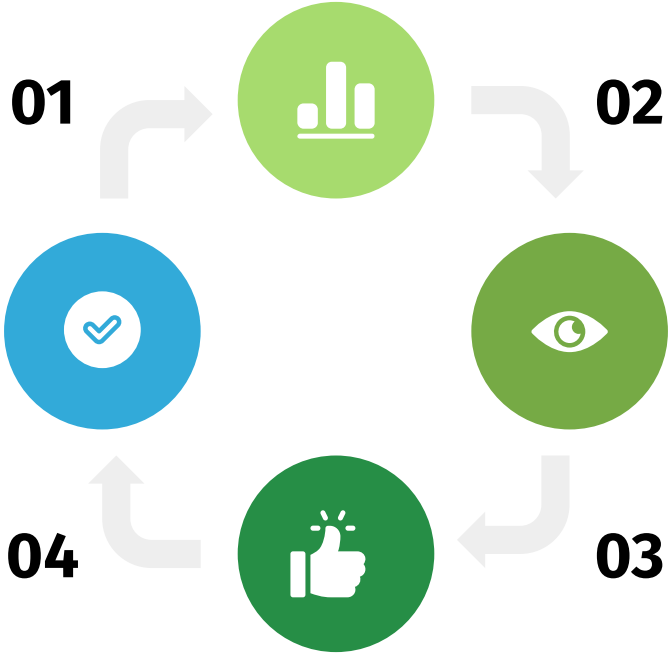
**Energy Indicators**

**Carbon Emission**

# Let’s Vizz Through Our Data....



# Modeling Path



01 Feature Cleaning & Transformation

02 Model Selection

03 Model Fitting & Prediction

04 Model Evaluation

# Feature Cleaning & Transformation

- **Handling missing data:** Features with less than 75% missing values were kept.
- **The feature transformations include:** Taking logarithm of highly skewed variables to make them normally distributed, Standardization, One-hot encoding.
- **Added new features:** Quadratic and interaction terms of continuous variables.

# Model Selection

MultiLinear Regression Model

Ridge Model

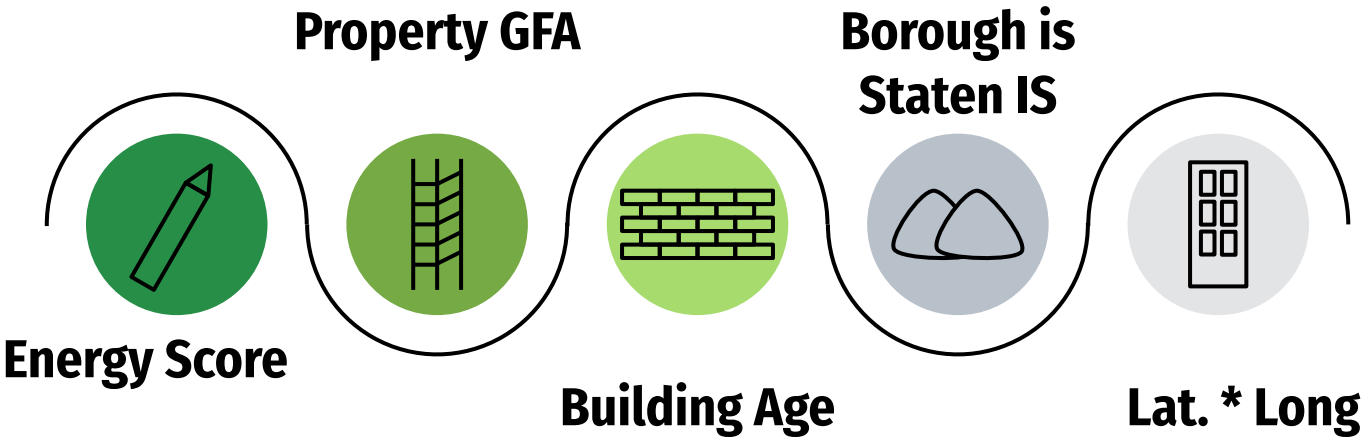
LASSO Model

Random Forest Model

# Model Evaluation

| | OLS | Ridge | LASSO | Random Forest |
|---|---|---|---|---|
| Train Set (0.9) | 0.60 | 0.59 | 0.58 | **0.73** |
| Test Set (0.1) | 0.58 | 0.58 | 0.58 | **0.68** |

# Top 5 Features According to Random Forest Model



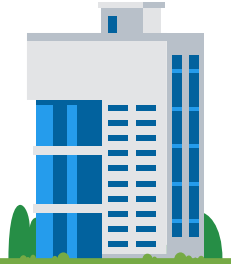
# Model Inference

- Inference was performed based on a linear model and alpha of **0.01**
- Standard errors were estimated, t-statistics and p-values were calculated

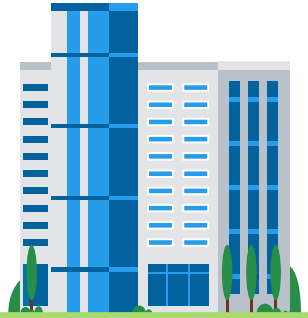
| Feature | Unit | Coefficient | p-value |
|---|---|---|---|
| **Building Age** | kBtu/SqFt/Year | 80.2 | 0.001 |
| **Property GFA** | kBtu/SqFt/SqFt | 39.4 | 0.24 |
| **Borough is Staten IS** | kBtu/SqFt | 30.44 | 3e-5 |
| **Property GFA * Longitude** | kBtu/SqFt.Degree | 89.9 | 0.006 |

# Recommendations Based on Inference

**Upgrade Old Buildings**



**Collect Data on All Buildings**



**Use Energy From Green Sources**

# Conclusion

- NYC buildings data is a rich and publicly available dataset that can be used to explore energy consumption patterns and answer important policy/research questions.

- Using a Random Forest model achieved the test $R^2$ of 0.68 which was the highest score value among the other models.

- From linear regression model, statistically significant features in explaining building energy consumption (such as building age) were identified and some recommendations were provided.

- For future work, by using data from multiple years and linking this dataset with other available data sources, we can build improved models.