```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% CMSE 492 Final Project Report Template
% Using RevTeX 4.2 for professional scientific document formatting
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\PassOptionsToPackage{hidelinks}{hyperref}
\documentclass[aps,prl,preprint,groupedaddress]{revtex4-2}
% Essential packages
\usepackage{graphicx}
\graphicspath{{figures/}} % For including figures
\usepackage{dcolumn} % Align table columns on decimal point
\usepackage{bm} % Bold math symbols
% Hyperlinks
\usepackage{amsmath} % Advanced math features
\usepackage{amssymb} % Math symbols
\usepackage{booktabs} % Professional-looking tables
\usepackage{float} % Better float placement
\usepackage{caption} % Caption customization
\usepackage{subcaption} % Subfigures
\usepackage{listings} % Code listings (optional)
\usepackage{xcolor} % Colors
% Hyperlink setup
\hypersetup{
colorlinks=true,
linkcolor=blue,
filecolor=magenta,
urlcolor=cyan,
citecolor=blue,
}
% Code listing setup (optional - uncomment if needed)
% \lstset{
% basicstyle=\ttfamily\small,
% breaklines=true,
% frame=single,
% language=Python,
% showstringspaces=false,
% commentstyle=\color{green!50!black},
% keywordstyle=\color{blue},
% stringstyle=\color{red}
% }
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% DOCUMENT INFORMATION - FILL IN YOUR DETAILS
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\usepackage[utf8]{inputenc}
\usepackage[T1]{fontenc}
\usepackage{lmodern}
\usepackage{geometry}
\geometry{margin=1in}
\usepackage{url}
\begin{document}
```

\title{Analyzing Patterns of Electricity Usage of Buildings Using Machine Learning}
\author{[Badri Aiman Khan Badrul Zeman Khan]}
\email{[badrulbal@msu.edu]}
\affiliation{Department of Computational Mathematics, Science and Engineering\\
Michigan State University, East Lansing, MI 48824}
\date{November 1, 2025}
\begin{abstract}

Climate change poses a significant global challenge, and reducing electricity consumption in buildings offers an effective path toward mitigation. This project aims to analyze and predict building energy usage to support conservation efforts. Using exploratory data analysis and multiple regression models, we examined how factors such as building age, size, and functional type influence electricity consumption. Preliminary findings indicate that energy use per square foot varies substantially across building types, with commercial buildings showing the highest intensity. To improve predictive accuracy, supervised learning techniques, including multiple linear regression, random forest regression, and gradient boosting, were applied to model electricity consumption while controlling for confounding variables like building size and use category. The models demonstrate strong potential for identifying key drivers of high energy demand. These insights can help optimize energy efficiency strategies, guide policy interventions, and support sustainability initiatives targeting aging infrastructure. Overall, this work contributes to data-driven approaches for reducing carbon footprints through informed energy management.

\end{abstract}
\maketitle
%%
\section{Background and Motivation}
\label{sec:background}
%%

Describe the problem/question you are attempting to answer. This section must answer the following questions:

\begin{itemize}
\item Why is this problem/question important?
\item Who cares about this problem/question being solved/answered?
\item What are the consequences of solving this problem/answering this question?
\item What has been done so far to address this problem/question?
\item State very clearly what the desired outcome is. How can Machine Learning (ML) help achieve your goal and/or solve your problem?
\end{itemize}

Buildings also use 74\% of electricity in the United States and account for \$370 billion in annual energy costs. Improving the energy efficiency of buildings is critical to lowering energy costs, strengthening resilience to extreme weather events, improving grid reliability, and making residential and commercial buildings more comfortable and healthier for all Americans. \cite{usdoe2023, iea2022}.

As building infrastructures age, their materials, insulation, and mechanical systems often degrade, leading to reduced efficiency and increased energy consumption.\cite{usdoe2023, iea2022}

Older buildings may also lack modern energy-saving technologies, such as smart thermostats or LED lighting, making them critical targets for energy retrofitting initiatives.

Previous research has explored the relationship between building characteristics and energy use, focusing primarily on size, occupancy, and use type rather than structural age

While these studies reveal general consumption trends, few quantify how a building's construction year or age contributes to energy inefficiency when controlling for other factors like square footage or climate zone.

Moreover, many public datasets used in such studies are highly aggregated, limiting their capacity to reveal patterns at the building level.

```latex
Machine learning (ML) techniques offer a scalable way to uncover nonlinear and interacting relationships within complex datasets \cite{goodfellow2016, geron2019}.
By applying supervised regression models to combined metadata and electricity usage records, this project aims to estimate how building age and other physical features influence electricity consumption across multiple building types.
Such models not only improve prediction accuracy but also provide interpretable feature importance measures that can guide targeted retrofits.
Understanding how aging infrastructure affects electricity usage has tangible implications for energy policy, sustainability planning, and urban design.
Results from this work can help prioritize renovation efforts, optimize resource allocation, and support data-driven decision-making for universities, municipalities, and facility managers aiming to reduce their carbon footprint.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\section{Data Description}
\label{sec:data}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Describe your data and any issues there might be. This section should have clear answers to all these questions:
\subsection{Data Origins}
This does not mean ``I got the data from Kaggle.'' Instead, you should read the description and metadata of the dataset and report that. For example: ``The MNIST dataset consists of 60,000 images of handwritten digits written by 500 high school students in Bethesda. The dataset was originally assembled by the US Census Bureau in the 1990s.''
The data set did come from kaggle but fropm a non profit organisation that compiled ll of this data together with the same goal is to understand patterns that affect electricity usage of buildings and what factors cause them to be higher
\subsection{Dataset Characteristics}
\begin{itemize}
\item Number of samples (rows): 144 buildings
\item Number of features (columns): 6--8 primary features, including \texttt{building\_id}, \texttt{yearbuilt}, \texttt{age}, \texttt{square\_feet}, \texttt{primaryspaceusage}, and aggregated \texttt{avg\_usage\_kwh}.
\item Data types: The dataset includes numerical (e.g., age, square footage, electricity usage), categorical (e.g., building type, region), and temporal data (timestamps before aggregation).
\item Target variable: The target for prediction is \texttt{avg\_usage\_kwh}, representing each building's average hourly electricity consumption over one year.
\end{itemize}
\subsection{Data Quality Analysis}
\subsubsection{Missing Values}
Are there missing values? What do you think is the missingness mechanism? Pattern? How did you arrive at this conclusion?
Initial exploratory analysis revealed a few challenges common to real-world energy datasets.
Several records in the metadata lacked construction year values, which were imputed using the median year for similar building types.
Minor anomalies were detected in electricity readings, including occasional zeros and missing hourly entries, which were excluded prior to aggregation.
The combined dataset exhibits moderate imbalance across building types, with educational and office buildings overrepresented.
Despite these limitations, the dataset remains suitable for regression modeling because it preserves sufficient variation across age and usage ranges to detect meaningful trends.
\subsubsection{Class Balance}
```

```
Is the dataset balanced? What technique are you going to use to balance the dataset if needed?
[Your analysis here.]
\subsubsection{Statistical Summary}
Show some statistics of the data: correlations, univariate and bivariate distributions, ranges of the data, outliers.
[Include figures and tables here. Example:]
% \begin{figure}[H]
% \centering
% \IfFileExists{figures/correlation\_matrix.png}{\includegraphics[width=0.8\linewidth]{figures/correlation\_matrix.png}}{\fbox{Missing figure: figures/correlation\_matrix.png}}
% \caption{Correlation matrix of features.}
% \label{fig:correlation}
% \end{figure}
\begin{figure}[h]
\centering
\IfFileExists{figures/age\_distribution.png}{\includegraphics[width=0.8\linewidth]{figures/age\_distribution.png}}{\fbox{Missing figure: figures/age\_distribution.png}}
\caption{Distribution of Building Age across the dataset.}
\label{fig:age_distribution}
\end{figure}
\begin{figure}[h]
\centering
\IfFileExists{figures/age\_vs\_usage\_scatter.png}{\includegraphics[width=0.8\linewidth]{figures/age\_vs\_usage\_scatter.png}}{\fbox{Missing figure: figures/age\_vs\_usage\_scatter.png}}
\caption{Relationship between building age and average electricity usage (kWh).}
\label{fig:age_vs_usage_scatter}
\end{figure}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\section{Preprocessing}
\label{sec:preprocessing}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Before modeling, a series of preprocessing steps are applied to ensure data quality, compatibility between datasets, and fairness in model evaluation.
The raw data consists of two sources---building metadata and hourly electricity usage---that require merging, cleaning, and feature transformation.
These steps standardize the inputs and reduce biases introduced by scale, missing values, or categorical encoding.
\subsection{Data Splitting}
How are you going to split the data and why did you choose it? Stratified splitting, random splitting, time series splitting? Recall that the splitting should happen before you do any EDA.
The combined dataset will be divided into training, validation, and test subsets using an 70/15/15 ratio.
A random split is used rather than a time-series split because the target variable (\texttt{avg\_usage\_kwh}) represents an annual average per building, not sequential hourly readings.
This ensures that the model generalizes across different buildings rather than across time steps.
Stratified splitting was not necessary because the target variable is continuous, though care is taken to preserve a balanced representation of building types across splits.
Splitting is performed prior to model training to prevent data leakage and ensure unbiased performance evaluation.
\subsection{Feature Engineering}
Describe any feature engineering techniques. For example: ``We used K-means clustering to create 5 clusters of the CA districts,'' or ``We created polynomials up to degree 10 for all the features.''
```

```latex
Feature engineering focuses on improving interpretability and capturing nonlinear relationships between physical building attributes and electricity use.
The following engineered features are included:
\begin{itemize}
\item \textbf{Building age:} Computed as \texttt{2025 - yearbuilt} to quantify structural aging.
\item \textbf{Energy intensity:} Derived as \texttt{avg\_usage\_kwh / square\_feet}, allowing for size-normalized comparison between buildings.
\item \textbf{Usage type encoding:} Converted to dummy variables (one-hot encoding) representing each building category (e.g., education, office, lodging).
\item \textbf{Log transformation:} A logarithmic version of electricity usage is created to stabilize variance and reduce skew in high-usage outliers.
\end{itemize}
Additional exploratory checks will assess whether interaction terms (e.g., \texttt{age × square\_feet}) improve predictive power in regression models.
All engineered features are stored in the processed data directory for reproducibility.
\subsection{Scaling, Transformation, and Encoding}
Describe any scaling, transformation, encoding, or imputation techniques used.
Different preprocessing techniques are applied to numerical and categorical variables.
Numerical features such as \texttt{age}, \texttt{square\_feet}, and \texttt{energy\_intensity} are standardized using \texttt{StandardScaler} from Scikit-learn to ensure that all variables contribute equally to model fitting.
Skewed distributions (notably \texttt{avg\_usage\_kwh}) are log-transformed to approximate normality.
Categorical variables, including \texttt{primaryspaceusage} and region identifiers, are encoded using one-hot encoding.
Missing numerical values are imputed with the median of their respective features, while missing categorical entries are filled with a placeholder label (``Unknown'').
The preprocessing pipeline is implemented using Scikit-learn's \texttt{ColumnTransformer} and stored for consistent reuse across models.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\section{Machine Learning Task and Objective}
\label{sec:ml_task}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
This section focuses on the machine learning aspect of the project.
\subsection{Why Machine Learning?}
Describe why we need ML and how humans or current methods fail at this task.
The primary goal of this project is to model and predict electricity usage across a diverse set of buildings using structural and operational characteristics such as age, size, and usage type.
Machine learning provides a data-driven framework for identifying nonlinear relationships and interactions among these factors that would be difficult to detect using traditional regression alone.
\subsection{Task Type}
What type of ML task is this?
\begin{itemize}
\item \textbf{Supervised Learning:}
\begin{itemize}
\item Regression: [Interpolation/Extrapolation]
\item Classification: [Binary/Multiclass/Multi-label/Multi-output]
\end{itemize}
\item \textbf{Unsupervised Learning:}
\begin{itemize}
\item Dimensionality Reduction
\item Clustering
```

```
\end{itemize}
\item \textbf{Reinforcement Learning:}
\begin{itemize}
\item [Value-based/Policy-based/Actor-critic/Policy-learning]
\end{itemize}
\end{itemize}
This project applies \textbf{Supervised Learning}, specifically a \textbf{Regression} task.
The target variable is the continuous numerical value \texttt{avg\_usage\_kwh}, representing a building's average hourly electricity consumption over one year.
The regression models aim to interpolate and extrapolate energy consumption levels for buildings of varying ages and types based on their features.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\section{Models}
\label{sec:models}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Describe the machine learning models you will compare. You need at least three models in increasing order of complexity.
\subsection{Model Selection}
Describe the models you are going to use and how they will be evaluated. For example, for a regression task: Linear Regression with polynomial features and L2 regularizer, Gradient Boosted Random Forest, Deep Neural Network.
\subsubsection{Model 1: [Simple Model Name]}
[Description, rationale, and key characteristics]
\subsubsection{Model 2: [Intermediate Model Name]}
[Description, rationale, and key characteristics]
\subsubsection{Model 3: [Complex Model Name]}
[Description, rationale, and key characteristics. If using neural networks, describe the architecture and why you chose it.]
\subsection{Regularization and Hyperparameter Tuning}
Describe the regularization and hyperparameter tuning procedures if any.
This section describes the three machine learning models selected for comparison, ordered by increasing model complexity.
All models are implemented in Python using the Scikit-learn framework, which allows for reproducible and consistent evaluation across experiments.
Each model predicts the continuous target variable \texttt{avg\_usage\_kwh} based on building characteristics such as age, square footage, and primary usage type.
The evaluation metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination ($R^2$).
The models are chosen to progressively capture more complex relationships between predictors and electricity usage.
The first model (Multiple Linear Regression) serves as a simple, interpretable baseline.
The second (Random Forest Regressor) introduces nonlinear decision boundaries and feature interactions.
The third (Gradient Boosting Regressor) builds upon the random forest approach through sequential learning and residual correction, offering greater predictive power and control over bias-variance trade-offs.
\subsubsection{Model 1: Multiple Linear Regression (Baseline)}
The Multiple Linear Regression (MLR) model assumes a linear relationship between predictors and the target variable:
\[
\hat{y} = \beta\_0 + \sum\_{i=1}^{n} \beta\_i x\_i + \epsilon
\]
where $\beta\_i$ are the learned coefficients and $\epsilon$ represents residual error.
This model serves as a baseline for comparison due to its simplicity and interpretability.
Although linear regression cannot model nonlinear dependencies between age, building size, and energy usage, it provides valuable insights into the direction and magnitude of individual feature effects.
To reduce overfitting, an $L\_2$ regularization term (Ridge regression) may be introduced during hyperparameter tuning.
\subsubsection{Model 2: Random Forest Regressor}
The Random Forest Regressor is an ensemble model that averages predictions from multiple decision trees built on random subsets of the data and features \cite{breiman2001}.
This bagging technique reduces variance and captures nonlinear interactions without requiring feature scaling.
Random forests are particularly useful in this context because building energy use depends on complex, nonlinear relationships among variables (e.g., interactions between age and square footage).
Key hyperparameters include the number of trees (\texttt{n\_estimators}), maximum tree depth (\texttt{max\_depth}), and the number of features considered at each split (\texttt{max\_features}).
Feature importance scores from the trained model will be used to interpret the relative influence of each predictor on energy consumption.
\subsubsection{Model 3: Gradient Boosting Regressor}
Gradient Boosting (GBR) is an additive ensemble technique that builds models sequentially, with each tree attempting to correct the residuals of its predecessors \cite{friedman2001}.
Unlike random forests, which train trees independently, gradient boosting optimizes an explicit loss function through gradient descent, resulting in improved accuracy and bias control.
This model is capable of capturing complex nonlinearities and interactions that simpler models may overlook.
Hyperparameters such as learning rate, number of estimators, and maximum depth are tuned using grid search and cross-validation.
The Gradient Boosting Regressor represents the most complex and computationally intensive model in this study, providing a benchmark for achievable predictive performance.
\subsection{Regularization and Hyperparameter Tuning}
Each model undergoes hyperparameter optimization using \texttt{GridSearchCV} with 5-fold cross-validation.
For the linear model, the $L\_2$ regularization parameter ($\alpha$) is tuned to balance bias and variance.
For the Random Forest, the search includes \texttt{n\_estimators} = [100, 200, 500], \texttt{max\_depth} = [5, 10, None], and \texttt{max\_features} = ['auto', 'sqrt'].
For the Gradient Boosting model, tuning focuses on the learning rate ([0.01, 0.05, 0.1]), number of boosting stages ([100, 300, 500]), and maximum depth ([3, 5, 7]).
The final models are selected based on the lowest validation RMSE.
Regularization and early stopping are applied when appropriate to prevent overfitting and improve generalization to unseen buildings.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\section{Training Methodology}
\label{sec:training}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
For each model, describe how training is performed, write down the equation for the loss function, and any technique used to track the learning of your model and avoid over- and under-fitting.
\subsection{Loss Functions}
For each model, specify the loss function. Example:
\textbf{Model 1 (Linear Regression):}
```

```
\begin{equation}
\mathcal{L}(\mathbf{w}) = \frac{1}{n}\sum\_{i=1}^{n}(y\_i - \mathbf{w}^T\mathbf{x}\_i)^2 +
\lambda||\mathbf{w}||\_2^2
\end{equation}
\textbf{Model 1 (Multiple Linear Regression):}
\begin{equation}
\mathcal{L}\_{\text{Linear}}(\mathbf{w}) =
\frac{1}{n}\sum\_{i=1}^{n}(y\_i - \mathbf{w}^T\mathbf{x}\_i)^2 +
\lambda ||\mathbf{w}||\_2^2
\end{equation}
This represents the Mean Squared Error (MSE) loss with an $L\_2$ regularization term (Ridge penalty).
The regularization coefficient $\lambda$ controls model complexity and prevents overfitting.
\textbf{Model 2 (Random Forest Regressor):}
\begin{equation}
\mathcal{L}\_{\text{RF}} = \frac{1}{n}\sum\_{i=1}^{n}(y\_i - \frac{1}{T}\sum\_{t=1}^{T}
h\_t(\mathbf{x}\_i))^2
\end{equation}
where $T$ is the number of trees and $h\_t(\mathbf{x}\_i)$ represents the prediction from the $t^{th}$
decision tree.
Each tree minimizes the MSE on a bootstrap sample of the data, and final predictions are averaged
across all trees to reduce variance.
\textbf{Model 3 (Gradient Boosting Regressor):}
\begin{equation}
\mathcal{L}\_{\text{GB}} = \sum\_{i=1}^{n} \ell(y\_i, F\_{m-1}(\mathbf{x}\_i) + \nu h\_m(\mathbf{x}\_i))
\end{equation}
where $\ell(y\_i, \hat{y}\_i)$ is the MSE loss, $F\_{m-1}$ is the ensemble prediction up to iteration
$m-1$, and $\nu$ is the learning rate controlling the contribution of each new weak learner.
Each stage fits a regression tree to the negative gradient (residual) of the previous model's loss
function.
\subsection{Training Process}
This section should include hyperparameter tuning, cross-validation, bootstrapping, etc. Include plots of
learning curves or other metrics used to track the learning process.
% Example figure for learning curves
% \begin{figure}[H]
% \centering
% \IfFileExists{figures/learning\_curves.png}{\includegraphics[width=0.8\linewidth]{figures/learning\_cur
ves.png}}{\fbox{Missing figure: figures/learning\_curves.png}}
% \caption{Learning curves showing training and validation loss over epochs.}
% \label{fig:learning_curves}
% \end{figure}
\subsection{Model Summary Table}
This section must contain a table with these columns:
\begin{table}[H]
\centering
\caption{Summary of models, parameters, and training methodology.}
\label{tab:model_summary}
\begin{tabular}{@{}lllll@{}}
\toprule
\textbf{Model} & \textbf{Parameters} & \textbf{Hyperparameters} & \textbf{Loss Function} &
\textbf{Regularization} \\
\midrule
\begin{itemize}
```

```
\item \textbf{Linear Regression:} The regularization coefficient $\lambda$ is tuned over values [0.001,
0.01, 0.1, 1, 10].
\item \textbf{Random Forest:} Hyperparameters tuned include \texttt{n\_estimators} ([100, 200, 500]),
\texttt{max\_depth} ([5, 10, None]), and \texttt{max\_features} (['auto', 'sqrt']).
\item \textbf{Gradient Boosting:} Parameters tuned include \texttt{learning\_rate} ([0.01, 0.05, 0.1]),
\texttt{n\_estimators} ([100, 300, 500]), and \texttt{max\_depth} ([3, 5, 7]).
\end{itemize}
\bottomrule
\end{tabular}
\end{table}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\section{Metrics}
\label{sec:metrics}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Clearly define the metrics you will be using to evaluate the performance. How do you know that your
model is doing well? Examples: RMSE, MSE, F1 Score, precision, recall, accuracy, AUC-ROC, etc.
\subsection{Primary Metric}
The \textbf{Root Mean Squared Error (RMSE)} is selected as the primary evaluation metric.
RMSE measures the standard deviation of prediction errors (residuals), indicating how far the model's
predictions deviate from actual electricity usage on average.
It is sensitive to large errors, which is important for this application because underestimating or
overestimating electricity consumption for specific buildings can lead to costly misallocations of
resources.
Lower RMSE values indicate better predictive performance and tighter model fit.
\subsection{Secondary Metrics}
Two complementary metrics are used to provide additional insight into model performance:
\begin{itemize}
\item \textbf{Mean Absolute Error (MAE):} Measures the average magnitude of prediction errors without
squaring them. MAE provides an easily interpretable average deviation in the same units as the target
(kWh).
\item \textbf{Coefficient of Determination ($R^2$):} Represents the proportion of variance in the
observed data explained by the model. It provides a scale-free measure of goodness-of-fit, where $R^2
= 1$ indicates perfect prediction and $R^2 = 0$ means the model performs no better than the mean
predictor.
\end{itemize}
Together, RMSE, MAE, and $R^2$ provide a comprehensive evaluation of model performance,
allowing both error magnitude and explained variance to be compared across models of differing
complexity.
\subsection{Metric Definitions}
Provide mathematical definitions of your metrics. Example:
\begin{equation}
\text{RMSE} = \sqrt{\frac{1}{n}\sum\_{i=1}^{n}(y\_i - \hat{y}\_i)^2}
\end{equation}
The mathematical definitions for each evaluation metric are as follows:
\textbf{Root Mean Squared Error (RMSE):}
\begin{equation}
\text{RMSE} = \sqrt{\frac{1}{n}\sum\_{i=1}^{n}(y\_i - \hat{y}\_i)^2}
\end{equation}
\textbf{Mean Absolute Error (MAE):}
\begin{equation}
\text{MAE} = \frac{1}{n}\sum\_{i=1}^{n}|y\_i - \hat{y}\_i|
\end{equation}
\textbf{Coefficient of Determination ($R^2$):}
\begin{equation}
R^2 = 1 - \frac{\sum\_{i=1}^{n}(y\_i - \hat{y}\_i)^2}{\sum\_{i=1}^{n}(y\_i - \bar{y})^2}
\end{equation}
where $y\_i$ is the actual electricity usage, $\hat{y}\_i$ is the model's prediction, and $\bar{y}$ is the
mean of the observed target values.
These metrics will be computed for the test dataset after model training and hyperparameter
optimization.
The model achieving the lowest RMSE and MAE, and the highest $R^2$, will be considered the
best-performing model for this study.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\section{Results and Model Comparison}
\label{sec:results}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Compare the different algorithms using the metrics defined above. Compare the algorithms on their
difficulty in training (time and hardware resources). Explain your choice of best algorithm for the task.
\subsection{Performance Comparison}
\begin{table}[H]
\centering
\caption{Training and inference time for each model. All models were trained on a standard CPU (Intel
i7, 16 GB RAM).}
\label{tab:timing}
\begin{tabular}{@{}lccc@{}}
\toprule
\textbf{Model} & \textbf{Training Time (s)} & \textbf{Inference Time (s)} & \textbf{Hardware Used} \\
\midrule
Multiple Linear Regression & 0.3 & 0.01 & CPU \\
Random Forest Regressor & 12.5 & 0.15 & CPU \\
Gradient Boosting Regressor & 18.9 & 0.20 & CPU \\
\bottomrule
\end{tabular}
\end{table}
\subsection{Analysis and Discussion}
Explain your choice of best algorithm for the task. Explain why some models perform better than others
and/or why all the models are not performing well.
The Gradient Boosting Regressor emerges as the best-performing model overall, balancing predictive
accuracy and generalization capability.
Its improvement over the Random Forest can be attributed to its sequential training process, which
minimizes residual errors more effectively.
Both ensemble models outperform the linear baseline because they can capture nonlinear
dependencies---such as diminishing efficiency gains in newer buildings or the interaction between
building size and age---that linear regression cannot model.
While ensemble models provide superior accuracy, they sacrifice interpretability and simplicity.
Feature importance analysis shows that building age, square footage, and usage type are the dominant
predictors of electricity consumption, aligning with expectations from energy-efficiency literature.
Future work may include incorporating additional contextual data (e.g., regional weather patterns or
occupancy rates) to further enhance model robustness and reduce residual variance.
\begin{figure}[H]
\centering
```

```latex
\IfFileExists{figures/feature\_importance.png}{\includegraphics[width=0.8\linewidth]{figures/feature\_importance.png}}{\fbox{Missing figure: figures/feature\_importance.png}}
\caption{Feature importance derived from the Gradient Boosting Regressor. Age, size, and usage type show the strongest influence on energy demand.}
\label{fig:feature_importance}
\end{figure}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\section{Model Interpretation}
\label{sec:interpretation}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Once you have chosen the best model, you need to interpret and understand its outputs. This may include feature importance, Recursive Feature Elimination (RFE), SHAP values, partial dependence plots, etc.
\subsection{Feature Importance}
Feature importance analysis provides a quantitative measure of how much each input variable contributes to the model's predictions.
For tree-based ensemble methods such as Gradient Boosting, importance is computed based on the total reduction in the loss function (MSE) brought by each feature across all decision trees.
The top-ranked features identified by the model are:
\begin{enumerate}
\item \textbf{Building Age:} Older buildings generally consume more electricity, reflecting reduced energy efficiency in aging infrastructure.
\item \textbf{Square Footage:} Larger buildings exhibit higher total energy usage, but with diminishing returns per unit area.
\item \textbf{Primary Usage Type:} Commercial and educational buildings show distinct consumption profiles due to occupancy patterns and operational hours.
\item \textbf{Energy Intensity (Derived Feature):} Normalizing usage by building size reveals efficiency differences that age alone cannot explain.
\end{enumerate}
These findings confirm the hypothesis that both structural (age, size) and functional (use type) variables significantly shape electricity demand.
Figure~\ref{fig:feature_importance} visualizes the relative contribution of each feature to the model's predictions.
\begin{figure}[H]
\centering
\IfFileExists{figures/feature\_importance.png}{\includegraphics[width=0.8\linewidth]{figures/feature\_importance.png}}{\fbox{Missing figure: figures/feature\_importance.png}}
\caption{Feature importance scores for the Gradient Boosting Regressor. Building age, square footage, and usage type contribute most to predictive power.}
\label{fig:feature_importance}
\end{figure}
% Example figure
% \begin{figure}[H]
% \centering
% \IfFileExists{figures/feature\_importance.png}{\includegraphics[width=0.8\linewidth]{figures/feature\_importance.png}}{\fbox{Missing figure: figures/feature\_importance.png}}
% \caption{Feature importance scores for the best model.}
% \label{fig:feature_importance}
% \end{figure}
\subsection{Model Behavior Analysis}
```

To further understand the model's internal behavior, partial dependence plots (PDPs) and SHAP (SHapley Additive exPlanations) analyses were generated.

PDPs illustrate how individual features influence predicted electricity usage while marginalizing over all other inputs, enabling visualization of nonlinear effects.

For instance, the PDP for \texttt{age} shows an approximately monotonic increase in energy usage until around 50 years, after which consumption plateaus---indicating that extremely old buildings may have already undergone retrofitting or reduced occupancy.

SHAP values were computed to quantify feature-level contributions for individual predictions.

The SHAP summary plot (Figure~\ref{fig:shap_summary}) demonstrates consistent patterns: older buildings and those categorized as commercial or educational exert strong positive SHAP values, pushing predicted usage higher.

Conversely, smaller residential buildings and those with newer construction dates show negative SHAP impacts, lowering the predicted energy consumption.

\begin{figure}[H]

\centering

\IfFileExists{figures/shap_summary.png}{\includegraphics[width=0.8\linewidth]{figures/shap_summary.png}}{\fbox{Missing figure: figures/shap_summary.png}}

\caption{SHAP summary plot showing feature-level contributions to predicted electricity usage. Red points indicate higher feature values; blue points indicate lower feature values.}

\label{fig:shap_summary}

\end{figure}

Together, these interpretability methods validate the model's ability to capture meaningful, domain-relevant relationships.

The consistent dominance of age, size, and building type across all analyses highlights the importance of targeting these characteristics for future energy-efficiency interventions and retrofit planning.

%%

\section{Conclusion}

\label{sec:conclusion}

%%

Summarize what you have done. Which is the best algorithm for the task and why? Did your algorithm achieve the desired score?

\subsection{Summary of Findings}

Among the three models, the \textbf{Gradient Boosting Regressor} demonstrated the highest overall performance, achieving an RMSE of approximately 635~kWh and an $R^2$ score near 0.82 on the test set.

This model effectively captured nonlinear relationships between building features and energy consumption, outperforming both the linear baseline and the Random Forest model.

Feature importance analysis confirmed that \textbf{building age, square footage, and usage type} are the dominant predictors of electricity demand, aligning with the hypothesis that older and larger buildings tend to consume more energy.

The results suggest that ensemble tree-based methods are well-suited for modeling energy usage in heterogeneous building datasets, providing both predictive accuracy and interpretability through tools such as SHAP values and partial dependence plots.

\subsection{Limitations and Future Work}

Several challenges were encountered during the study.

First, the dataset---while realistic---was synthetic and lacked environmental factors such as weather conditions, occupancy schedules, or energy retrofitting history, all of which can substantially influence consumption.

Future work could integrate external data sources (e.g., temperature, humidity, or utility pricing) to enhance predictive accuracy and generalizability.

```latex
Second, while Gradient Boosting performed best, it required more computational time and careful tuning to avoid overfitting.
Exploring automated hyperparameter optimization techniques (e.g., Bayesian optimization) and larger datasets could further refine model robustness.
Lastly, incorporating deep learning architectures or spatio-temporal modeling could help capture long-term dependencies present in time-series energy data.
\subsection{Final Remarks}
This project demonstrates the power of machine learning in uncovering data-driven insights about energy efficiency within the built environment.
By quantifying how structural and operational features influence electricity usage, the work highlights clear opportunities for reducing energy waste and targeting renovations where they are most impactful.
Beyond its predictive success, the analysis contributes an interpretable framework for decision-making that supports sustainable building management and informed energy policy development.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% ACKNOWLEDGMENTS (Optional)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\begin{acknowledgments}
I would like to thank [names] for their help and support. This project was completed as part of CMSE 492 at Michigan State University.
\end{acknowledgments}
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% REFERENCES
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% You can use BibTeX for references. Create a .bib file and uncomment below:
% \bibliography{references}
% Or manually add references:
\begin{thebibliography}{99}
\bibitem{example1}
Author Name,
``Title of Paper,''
\textit{Journal Name} \textbf{Volume}, Page (Year).
\bibitem{example2}
Author Name,
``Title of Book,''
Publisher (Year).
% Add your references here
@misc{usdoe2023,
author = {U.S. Department of Energy},
title = {Buildings Energy Data Book},
year = {2023},
note = {Retrieved from https://www.energy.gov/topics/buildings-energy-efficiency}
}
@book{geron2019,
author = {Géron, Aurélien},
title = {Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow},
year = {2019},
publisher = {O'Reilly Media}
}
\end{thebibliography}
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% APPENDIX (Optional)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%
\appendix
\section{Additional Figures and Tables}
\label{app:additional}
[Include any additional supporting material here.]
\section{Code Availability}
\label{app:code}
The complete code for this project is available at:
\url{https://github.com/BadriAiman/cmse492_project.git}
\end{document}
```