ENSF 444 Final Project Presentation Script Title: Predictive Modeling for Household Energy Consumption Optimization

---

Slide 1: Title Slide Hello everyone, we're Group 50 from ENSF 444. Our final project centers around predictive modeling for household energy consumption optimization. This initiative, developed for our fictional client Green Leaf Energy, focuses on building a machine learning tool that empowers users to forecast and reduce their electricity use. We are excited to share how data science can help households lower costs, reduce environmental impact, and gain better control of their energy usage. Our model combines real-world data, rigorous ML approaches, and user-centric outcomes. Let's dive in and explore how our system transforms energy insights into sustainable action!

Slide 2&3: Project Motivation and Proposed Solution Household energy consumption is more critical than ever. With prices climbing and environmental stakes high, many homes are left without the tools needed to make smart decisions. Families are dealing with financial strain due to unpredictable energy bills. Even worse, the carbon emissions tied to unnecessary consumption are accelerating climate change. Without real-time forecasts or usage trends, users operate blindly—missing opportunities to optimize behavior and reduce waste. Green Leaf Energy identified this gap as a key opportunity for innovation.

Our solution is built around a robust machine learning model. By analyzing historical consumption patterns using the RECS 2020 dataset, we can train models to forecast future energy needs accurately. The system does more than just predict—it reveals which factors most influence a household's energy profile. Whether it's seasonal trends, appliance use, or occupancy, these insights empower users to act. The goal is practical: give people the foresight to plan smarter, use less, and save more—all while contributing to a greener grid.

Slide 4: Dataset Overview "We used the 2020 RECS dataset, which surveys U.S. residential energy use. Out of hundreds of variables, we selected 13 key features that influence energy use: such as square footage, heating type, number of rooms, insulation, and geographic location. The dataset included a mix of numeric and categorical variables."

Slide 5: Data Cleaning & Preprocessing "We replaced invalid survey responses marked as -1 and -2 with nulls and dropped rows with missing data. We categorized features into numerical and categorical groups. Numerical features were standardized using StandardScaler, while categorical ones were one-hot encoded for model compatibility."

Slide 6: Modeling Pipeline "We created a unified machine learning pipeline. This ensured consistent preprocessing before training any model. We split the data into 80% training and 20% testing. Three models were trained: Linear Regression, Decision

Tree Regressor, and XGBoost Regressor."

Slide 7: Model Performance "We evaluated models using MAE, RMSE, and R-squared. XGBoost outperformed others with an R-squared of 0.4982. Linear Regression performed reasonably, while the Decision Tree showed signs of overfitting. The plot on this slide shows XGBoost's predictions versus actual values."

Slide 8: Hyperparameter Tuning "To further optimize performance, we applied GridSearchCV on XGBoost. We tested multiple combinations of hyperparameters across 3-fold cross-validation. The best model achieved a slightly improved R-squared of 0.4985 with reduced overfitting."

Slide 9: Feature Engineering & Correlation "We engineered new features like square footage per bedroom and the log of energy usage. Our correlation heatmap confirmed strong relationships between total home square footage and energy use. These findings reinforced the importance of home size and insulation."

Slide 10: PCA and KMeans Visualization "To explore patterns visually, we used PCA to reduce dimensions and applied KMeans clustering. Households were grouped into three clusters with distinct energy behaviors. PCA helped simplify the dataset while preserving variance."

Slide 11: Cluster Insights "Cluster 1 had the largest, most energy-intensive homes with older equipment. Cluster 0 represented moderate households with fair insulation and warmth. Cluster 2 had smaller families, newer equipment, and the lowest indoor temperatures. Clustering helps design targeted energy-saving strategies."

Slide 12: Model Limitations "While our model performed well given the features, the R-squared score suggests that a large portion of energy use variation remains unexplained. This may be due to missing context like occupant behavior, local weather, or seasonality."

Slide 13: Future Improvements "To improve, we suggest adding external data such as climate records or utility rates, testing more advanced models like LightGBM, and exploring SHAP for explainability. Better features and deeper models could enhance predictive power."

Slide 14: Conclusion "To conclude, we built a robust machine learning pipeline that identifies energy use patterns in U.S. households. Our best model was XGBoost with optimized tuning. Cluster analysis provided additional insight. This solution supports Green Leaf Energy's mission to improve energy efficiency."