Determining Best Method to Predict Electricity Consumption in Dallas, Texas

CS 8321 Final Paper

Aitor Elfau
Computer Science
Southern Methodist University
Dallas TX
aelfaugonzalez@smu.edu

Sam Yao
Computer Science
Southern Methodist University
Dallas TX
samyao@smu.edu

Abdul Wasay
Computer Science
Southern Methodist University
Dallas TX
awasay@smu.edu

ABSTRACT

Extensive research on methods of predicting electricity production has been conducted using a multitude of machine learning and artificial intelligence techniques. This is important, because it has been shown that electricity load monitoring and predicting helps consumers lower their usage/costs and carbon footprint. However, the geographic epicenter for these research papers has mainly been the EU and Eastern Asian countries. We aim to synthesize datasets for the US market demographics and use them to predict electricity rates and usage with the same level of accuracy achieved by the other researchers.

KEYWORDS

Electricity Prediction, Linear Regression, Support Vector Regression, Extreme Gradient Boost, Random Forest, Wide and Deep Networks, US Electricity, DLM (Direct Load Monitoring), NILM (Non-Intrusive load monitoring), MSE (Mean Squared Error), R² (R-squared or the Coefficient of Determination)

1 Motivation

Research suggests that heightened awareness of electricity bills and comparative usage rates can lead to significant reductions in energy consumption. Trotta (2020) notes that when consumers are more informed about their energy expenses and how they stack up against others, they may reduce their utility usage by as much as 40%, with an average savings of 25% and a minimum of 10%. Similarly, Georgiou (2013) reports a substantial 23% decrease in electricity usage in Latvia. Furthermore, Fischer (2008) observed an average reduction of 12% across multiple studies, underscoring the potential for informed consumers to achieve meaningful energy savings. This awareness can be enhanced through social interaction and the real-time visualization of energy consumption, as highlighted by Mulvad (2012). Additionally, Trotta (2020) found that a significant proportion of people (two-thirds of respondents) are not fully aware of their electricity usage but are keen to learn more. Increased knowledge about energy consumption consistently leads to greater savings.

Efforts to monitor energy use through technology, such as the use of microcontrollers in Malaysia in 2020 (Malik, 2020), have been made. However, Bohdanowicz (2021) suggests that there is hesitancy among consumers to adopt new technologies for energy savings in their homes. Despite this, feedback mechanisms that provide detailed breakdowns of energy usage by appliance (Fischer, 2008) have proven effective, motivating our investigation into smarter, more user-friendly solutions for energy information dissemination.

The decision to avoid relying on self-reported data stems from its inherent inaccuracies and the need for extensive user participation or access to non-public data from power companies. Rashid et al. (2017) argue that people often lack a complete understanding of their electricity use, making them unreliable reporters for precise data analysis.

Our current research aims to identify the most effective AI and ML models for predicting electricity usage and comparing these predictions to user forecasts. Previous studies, such as Jin et al. (2021), have demonstrated the accuracy of AI and ML models in estimating energy use, however those studies mainly focused on the EU and China that has a vastly different climate than the state of Texas, where we want to deploy our model. These models are very particular with the type of appliances used (Coleman et al., 2012), the climate (Jones, 1992) as well as the type of houses/buildings available (Bennet, 2017). Hence, we wanted to test to find out the best ML/AI models for the American, specifically Texas, demographic data.

This platform is being designed to reduce monthly costs (Santarossa et al., 2016) and environmental impacts (Oar, 2015) through effective energy management. We aim to promote sustainable consumption behaviors through nudging, a method proven effective by Lehner et al. (2016).

In summary, our paper aims to answer the following research questions:

1. Researchers have investigated what is the best model for electricity prediction only *outside the United States*,

hence there is a lack of research done on the best type of ML/AI with the US specific demographics of weather / house size/ type of construction etc.

2. Research on how synthetic data being used to create a larger dataset within the context of electricity usage is not explored, hence, we want to see if it is possible to use raw datasets to come up with models that perform well, that can be useful enough to provide the initial small number of users meaningful insight into their electricity usage.

2 Related Works

Machine learning (ML) and artificial intelligence (AI) have been instrumental in advancing our understanding and management of energy data. Building upon existing research, we aim to refine these methods to achieve optimal accuracy for demographics specific to the southern United States. We have selected methodologies that have demonstrated efficacy on large datasets, high ($\mathbf{R^2}$) values, and practical applicability for rapid platform deployment.

Our approach includes looking over and deciding the best between the following models, which have shown promising results in various global contexts:

- Recurrent Neural Networks and Long Short-Term Memory Networks (modified RNNs) which achieved an ($\mathbf{R^2}$) of 0.96 in India, as reported by Chandrasekaran et al. (2023).
- A Stacking Model combining Random Forest, Gradient Boosted Decision Trees, Extreme Gradient Boosting, Support Vector Machines, and K-Nearest Neighbors, which achieved an (R^2) of 0.86 (Wang et al., 2020).
- An ensemble of a Three-Layer Feedforward Network, Radial Based Network, Adaptive Neuro-Fuzzy Inference System, and other neural networks, which resulted in an ($\mathbf{R^2}$) of 0.9840 during the colder periods in Norway, a critical time for heating energy consumption (Jovanović et al., 2015).
- A combination of Feed-Forward Back-Propagation Artificial Neural Network (ANN) with Random Forest (RF), which was tested in a single hotel in Madrid, Spain, and achieved an ($\mathbf{R^2}$) of 0.964 (Ahmad et al., 2017).
- Various models including Extreme Gradient Boosting (XGBoost), Random Forest (RF), Artificial Neural Network (ANN), Gradient Boosting Decision Tree (GBDT), and Support Vector Regression (SVR) applied in China, with the maximum ($\mathbf{R^2}$) reaching 0.6 (Wang et al., 2019).
- An extensive study comparing Linear Regression with Ordinary Least Squares (OLS), Random Forest (RF), Support Vector Regression (SVR), Multivariate Adaptive Regression Splines, Gaussian Process Regression (GPR), and Neural Networks (NN) across increasing sample sizes, achieving an ($\mathbf{R^2}$) of up to 0.99 (Østergård et al., 2018).
- A comprehensive review by Wei et al. (2018) that explores data-driven approaches in building energy analysis, focusing on both prediction methods like Artificial Neural Networks, Support Vector Machines, Statistical Regression, Decision Trees, and Genetic Algorithms, and classification techniques such as K-Means Clustering, Self-Organizing Maps, and Hierarchical Clustering.
- By integrating and adapting these advanced methodologies, we aim to develop a robust model that not only predicts energy usage accurately but also facilitates efficient platform rollout, specifically tailored for our target demographic in the southern United States. This strategic approach is designed to leverage cutting-edge technology to enhance energy management and promote sustainable practices.

3 Methods

We are developing a comprehensive model to analyze power utility bills collected through consumer self-reporting. This model will integrate several key variables, including demographic data, appliance profiles, and power consumption metrics. Our objective is to use this model to offer personalized advice to consumers on their electricity usage. For some, this will confirm that their consumption levels are satisfactory. For others, it will help identify potential issues such as malfunctioning appliances, misuse of energy, faulty meters, unoptimized tariff plans, or even illegal connections.

The goal of this project is to empower customers with better information about their energy consumption, enabling them to reduce their expenses and minimize wastage. This not only has financial benefits but also significant environmental implications.

Using our data sources, we will:

- Generate realistic synthetic data to supplement our observations as our dataset grows.
- Assess whether self-reported data might be inaccurate or anomalous when it significantly deviates from expected patterns.

This approach will not only improve the precision of our analysis but also enhance the overall effectiveness of the energy management solutions we provide to consumers. To evaluate the performance of the models in predicting values, we utilized two common statistical metrics: Mean Squared Error (MSE) and the $\mathbf{R^2}$ Coefficient.

Mean Squared Error (MSE): MSE measures the average squared difference between the estimated values and the actual value. It provides a way to quantify the error in prediction, with lower values indicating better model performance. MSE is particularly useful because it emphasizes larger errors due to the squaring of each term, which can be crucial for many practical applications where large errors are particularly undesirable (Chai & Daxter, 2014).

$\mathbf{R^2}$ Coefficient: This value provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. An ($\mathbf{R^2}$) value of 1 indicates perfect correlation between predicted and actual values, while a value of 0 indicates that the model does not improve prediction over using the mean of the actual values. In general, higher ($\mathbf{R^2}$) values indicate better model performance (Glantz & Slinker 1990).

To determine the nature of the Research Question, we trained five types of models: Linear Regression, Support Vector Regression (SVR), XGBoost, Random Forest, and a Wide and Deep Neural Network (which may be referred to as both a Wide and Deep

Network or simply as a Neural Network). Each of the models were trained on the dataset, with the data split 80/20 for training and validation respectively. We justify the 80/20 split because for our dataset, such a split would leave enough data for our models to train on while leaving a testing set big enough to validate our models (Douglass, 2020).

When applicable, GridSearchCV, a method for hyperparameter tuning, was used to identify optimal hyperparameters. The GridSearchCV function returns the most optimal set of hyperparameters input into the function for a particular type of model. This method was applied to one type of Linear Regression model, SVR, XGBoost, and Random Forest. We configured GridSearchCV to utilize the ‘Negative Mean Squared Error’ to determine which set of hyperparameters were the best to use. However, GridSearchCV does take a very long time to run, especially with the more parameter values tested, so in the case of the Ridge Regression sub model, guessing parameter values and fitting models based on those guesses was significantly faster. It should be noted that despite the use of both the R^2 and MSE scores as model metrics, only the Negative Mean Squared Error was used during hyperparameter tuning due to time constraints.

For the Neural Networks, we constructed a base neural network architecture consisting of an input layer with 115 neurons (equal to the number of columns in the dataset), followed by Dense layers of 64, 32, and 16 neurons, each with ReLu as their activation function (See Figures X and Y). Finally, the output layer consists of one neuron for the result, with ReLu as the activation function as well. ReLu was chosen as the activation function because the smallest value that can be output from ReLu is 0, which fits our use case because energy companies cannot charge customers negative prices. Variations were built upon this base architecture, consisting of removing middle layers or adding a Batch Normalization layer after the input layer. Another network was given pre-normalized data before training. A total of 10 combinations of neural networks were tested. GridSearchCV was not used on the Neural Networks.

To determine the metrics for our models, we utilized two methods to fit and train/test our models. The first method simply fit the model once, while the second method used KFold Validation. In the KFold Validation section, we utilized 10 folds per model. The average metric score (MSE and R^2) during training and testing were compiled, and the average score between the training and testing was used as the final metric.

3.1 Data

Our work leverages a structured dataset, which is segmented into several distinct subsets, each contributing vital information necessary for the comprehensive analysis of electricity consumption patterns:

- **Metadata Subsection:** This initial subset encompasses essential system data that aids in ensuring the overall reliability and validity of the collected information. Additionally, it includes contact details necessary for ongoing communication with participants, which is crucial for obtaining continuous data updates and verifying the accuracy of the information provided. To maintain the focus on actionable data and to ensure privacy and data integrity, this metadata is regularly purged following the validation process.

- **Demographic Information:** This segment of the dataset provides a detailed breakdown of the household demographics, capturing data points such as the age distribution of residents, the total number of occupants, and the duration of their residence at the property. These factors are instrumental in predicting variations in energy usage across different households. Initially, to simplify the demographic variables in our predictive models, we have amalgamated gender distinctions into a unified category, potentially to refine this in subsequent model iterations.

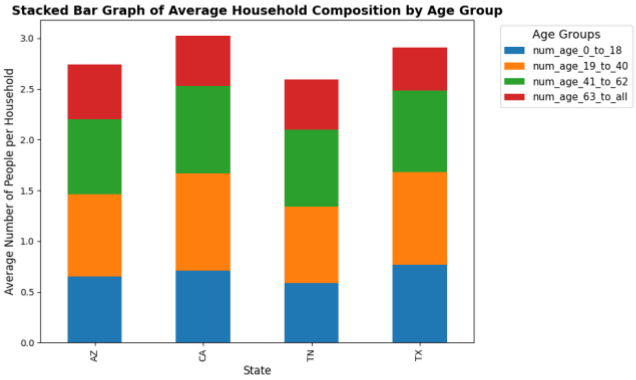


Figure 1: Average household composition. M. El Dayeh [29]

- **Dwelling Characteristics:** This subsection records comprehensive information about the physical attributes and location of the dwelling, including the square footage of the living space, the orientation relative to cardinal directions, and the construction materials used in the building's envelope, such as roofing, windows, and walls.

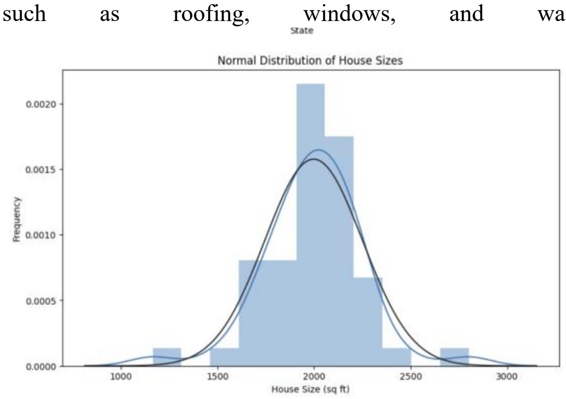


Figure 2: Normal distribution of house sizes G. Pandi [27]

Additionally, geographic positioning is documented to factor in regional climate impacts on energy consumption. This data is crucial for evaluating the energy efficiency potential of different residences and tailors our analysis to accommodate environmental variables.

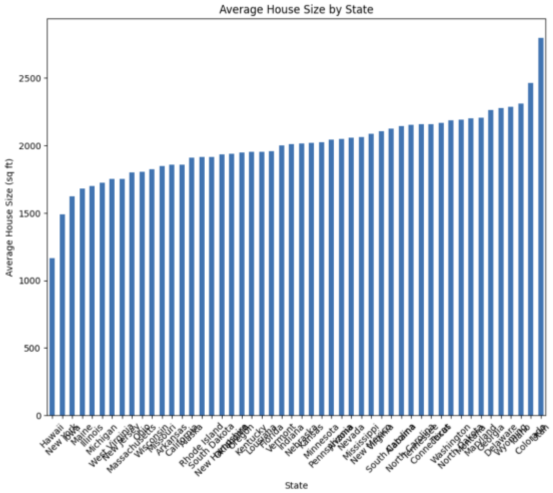


Figure 3: Average house size by state C. And [28]

- **Appliance Usage Metrics:** The primary focus of our analysis, this subset catalogs the quantity and types of appliances within each household, with particular attention to high-consumption devices such as electric heating and water heating systems. Understanding which appliances are used, along with their operational patterns, is essential for disaggregating total energy consumption into specific loads, thus allowing for more targeted energy-saving advice.
- **Consumption and Pricing Data:** At the core of our analysis lies the consumption data measured in kilowatt-hours (kWh). This metric serves as the universal standard across our models, enabling consistent comparison and regression analysis irrespective of local pricing variations. By uniformly applying kWh as the central measure, our approach allows for scalability and adaptability across different geographical regions and utility markets, not only within the state of Texas but potentially on an international scale. Additionally, we capture detailed pricing information corresponding to the consumption metrics. This pricing data is not only crucial for current cost analysis but also sets a foundation for future investigations into tariff structures and their impacts on consumer behavior. To ensure temporal accuracy and relevancy of our consumption and pricing data, we meticulously record the start and end times of the measurement periods. This temporal precision is essential for aligning our data with variable energy rates and consumption patterns, providing a robust framework for our predictive models. Through this comprehensive approach, we aim to equip consumers with precise, actionable advice based on their specific energy usage and the associated costs, ultimately fostering more informed and economically efficient energy consumption decisions.

We are generating synthetic data because no directly relevant datasets have been found. The data needed for this study is privately held by power companies for which it holds business value. And it is fractured into the different power companies. Furthermore, those datasets would only cover the consumption and not be split according to the demographics that are useful for this particular use case.

The option we would pursue in a future study data would be gathered via self-reporting by incentivizing consumers to register and provide this information in exchange for good faith advice and analysis. This analysis will be in part based on the models trained with their own data. Forming a positive feedback loop of better data and better guidance.

The isolated appliance consumption data can be obtained with two approaches:

- **Direct Load Monitoring (DLM)**: This method employs multiple individual measurement devices, each attached directly to major appliances within a household. This approach provides highly accurate and specific data on the energy consumption of each appliance by measuring their exact electrical usage in real-time. While DLM offers precision and granularity that can be advantageous for detailed energy audits and specific diagnostic applications, it also involves higher costs and greater complexity in terms of installation and maintenance. Each device must be individually installed and calibrated, which can be intrusive and disruptive for households. Additionally, the proliferation of devices increases the likelihood of maintenance issues and requires a more robust data management system to handle the increased volume of data.
- **Non-Intrusive Load Monitoring (NILM)**: Aligned with the nature of our study which stands out as a transformative tool in the field of energy management. NILM operates by analyzing variations in electrical load data captured at a single point, such as the household's main electrical panel, to deduce the operation and energy consumption of individual appliances within the home. This sophisticated analysis allows NILM systems to identify specific electrical signatures associated with different appliances, thereby disaggregating total energy consumption into specific end-uses without the need for individual appliance meters.

The full structure of the dataset can be found in the annex of this paper.

3.2 Linear Regression

Two types of Linear Regression models were tested: one model with no parameters or hyperparameter tuning, and another model used Ridge Regression, otherwise known as L2 Regularization. We chose L2 Regularization for its role in reducing overfitting, especially when there are many weights in the model. For the L2 model, only the alpha parameter was used, which was set to a value of 100.

3.3 Support Vector Regression (SVR)

This type of model was chosen due to the high dimensionality of the dataset. SVR works by creating hyperplanes between each observation in a way in which the distance between the observations and the hyperplanes are maximized. Two parameters were utilized: the kernel trick and the regularization parameter *C*. The kernel trick defines the shape of the hyperplanes, and the *C* parameter helps to control the margin of error between the observations and the hyperparameters. In this model, we used the Radial Basis Function (rbf) kernel, and a *C* value of 5000.

3.4 XGBoost

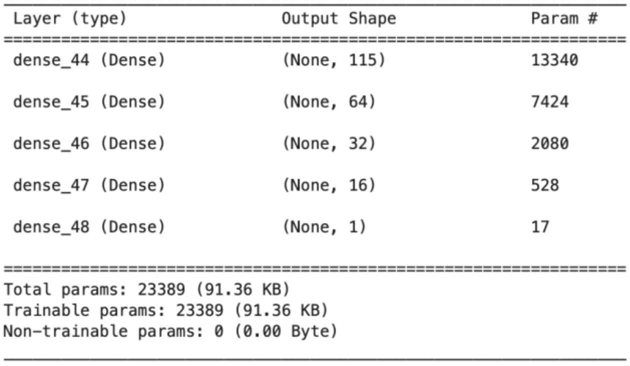
XGBoost is a decision tree algorithm that combines ensemble learning with gradient boosting. Like Random Forest in using ensemble learning, XGBoost goes beyond by training the models it selects from in parallel instead of iteratively. We utilized two parameters when constructing the XGBoost model: the Number of Estimators and the Maximum Depth of the Tree. We used a model with 8 Estimators and a Maximum Depth of 10.

3.5 Random Forest

Random Forest is a decision tree algorithm that employs ensemble learning. Similar to XGBoost, it trains a number of models and learns from them in order to improve the accuracy of the final decision tree. It also uses the same parameters as XGBoost (Number of Estimators and Maximum Depth of the Tree). This model uses 700 Estimators with a Maximum Depth of 4.

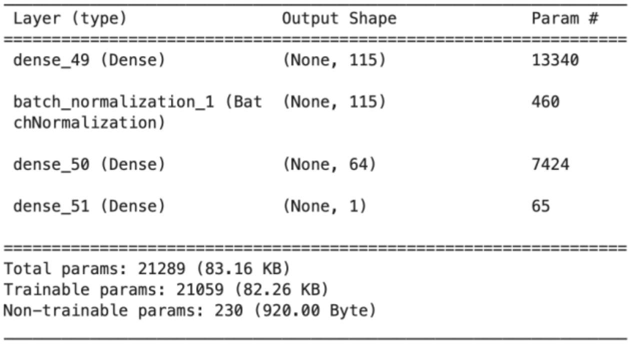
3.6 Wide and Deep Networks

As described earlier in this section, we trained and examined the performance of ten variations of a Wide and Deep Neural Network consisting of an input layer, three hidden layers, and an output layer. We chose three hidden layers to gradually Each model varied solely within the number and type of hidden layers. Only the input and output layers remained the same for all ten models. BatchNorm was added directly after the input layer in order to normalize the input. In one architecture, the dataset was normalized before being fed into the Network. We are using 10 different Network models in order to find least computationally taxing and best performing architecture for our use case.



```
Layer (type)                Output Shape              Param #
=================================================================
dense_44 (Dense)            (None, 115)               13340

dense_45 (Dense)            (None, 64)                7424

dense_46 (Dense)            (None, 32)                2080

dense_47 (Dense)            (None, 16)                528

dense_48 (Dense)            (None, 1)                 17

=================================================================
Total params: 23389 (91.36 KB)
Trainable params: 23389 (91.36 KB)
Non-trainable params: 0 (0.00 Byte)
```

Figure 4: Architecture of Neural Network Variant No.1



```
Layer (type)                Output Shape              Param #
=================================================================
dense_49 (Dense)            (None, 115)               13340

batch_normalization_1 (Bat  (None, 115)               460
chNormalization)

dense_50 (Dense)            (None, 64)                7424

dense_51 (Dense)            (None, 1)                 65

=================================================================
Total params: 21289 (83.16 KB)
Trainable params: 21059 (82.26 KB)
Non-trainable params: 230 (920.00 Byte)
```

Figure 5: Architecture of Neural Network Variant No. 9 with Batch Normalization

4 Results and Conclusion

When comparing the R^2 value between fitting the model once versus using KFold Validation, the R^2 Score does not vary significantly for each of the types of models, though the KFold Score is slightly higher than when the model is fit only once. Both Linear Regression models performed about the same, with plain Linear Regression scoring $R^2 = 0.829$ for fitting once and $R^2 = 0.834$ for KFold, while the L2 Regularized model scored $R^2 = 0.814$ for fitting once and scored $R^2 = 0.839$ during KFold. SVR Scored $R^2 = 0.858$ when fitted once and $R^2 = 0.871$ in KFold. XGBoost had the highest R^2 score overall, with $R^2 = 0.876$ when fitted once while $R^2 = 0.904$ when using KFold validation.

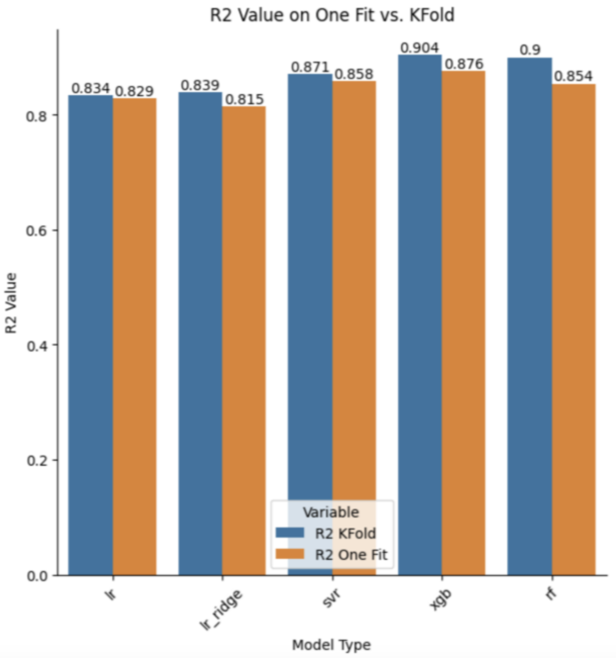


Figure 6: Comparison of R^2 values across single fitting vs. 10 KFold Validation

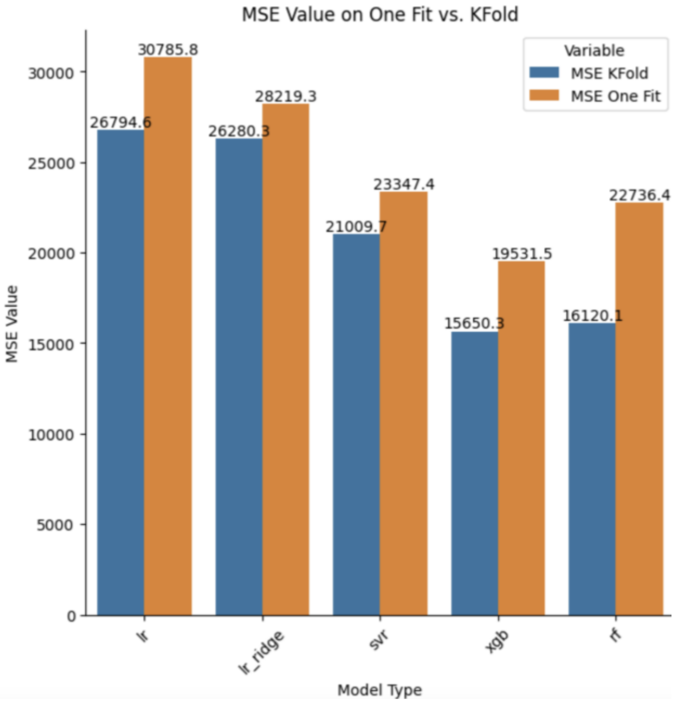


Figure 7: Comparison of MSE values across single fitting vs. 10 KFold Validation

However, the values between the training methods for MSE significantly differed. For plain Linear Regression, the MSE scores were MSE = 30,785.822 when the model was fitted once and MSE = 26,794.580 when utilizing KFolds. For L2 Regularization, MSE = 28,219.346 when fitted once in comparison to MSE = 26,280.328 when utilizing KFold. SVR scored MSE = 23,347.404 when fitted once, while scoring MSE = 21,009.671 when fitted using KFold. Random Forest scored MSE = 22,736.390 when fitted once and MSE = 16,120.143 when using KFold. XGBoost performed the best, scoring MSE = 19,531.461 when fitted once, while scoring MSE = 15,650.319 when using KFold.

Of the ten neural network architectures that were trained, only Neural Network Architectures 5, 6, 7, and 8 reached convergence. Architecture 1 failed to converge at all, scoring very low with R^2 as a performance metric. The remaining Architectures 2, 3, 4, 9, and 10 did seem to reach a point of convergence but then diverged as training progressed. Architectures 5, 6, 7, and 8 converged at an R^2 score of about 0.8, while converging at an MSE value of about 25000. Architecture 10 showed the lowest MSE loss values, which could be related to the data being normalized before training. The MSE converged to around 0.2 before diverging around the 30th epoch.

Architectures 5 through 8 had two or more hidden layers removed from the final design. Since these models were the only ones that converged, our results suggest that for this research question, a neural network with just one input and one output layer is more than enough for the task, and that any more hidden layers may result in overfitting. Normalizing the dataset before fitting the model could also improve training performance. However, the architectures tested did not beat the performance of the models tested previously, with models such as XGBoost and Random Forest scoring up to 0.10 higher for R^2 and around 10,000 lower for the MSE.

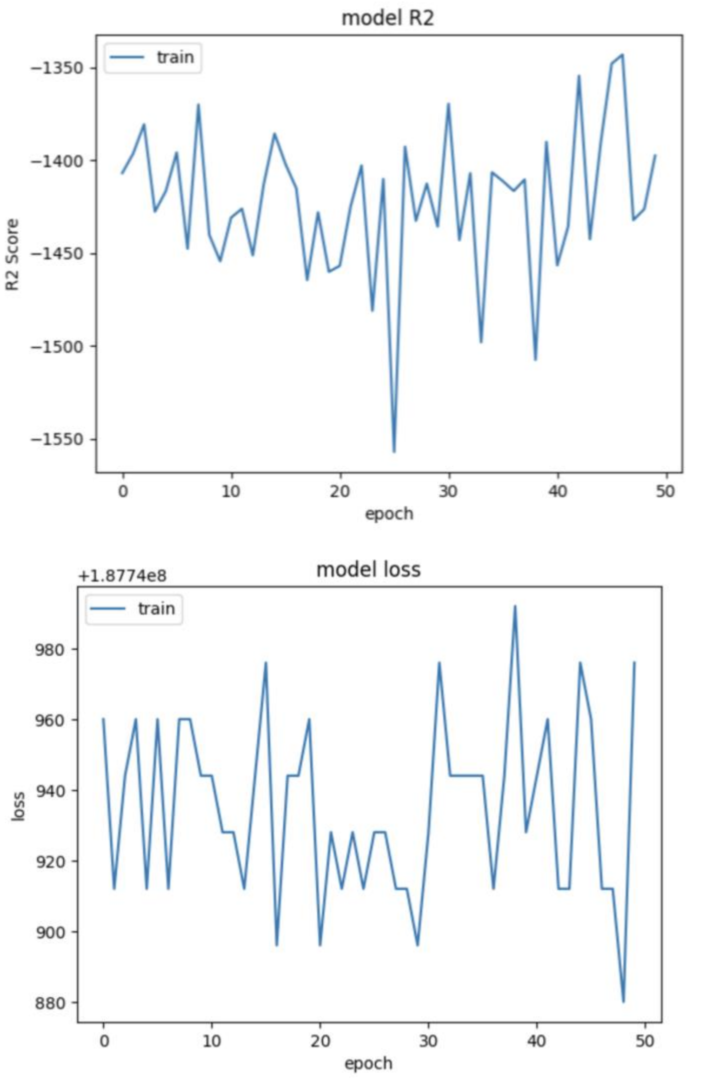


Figure 8: Training R^2 and MSE scores for Neural Network Architecture 1

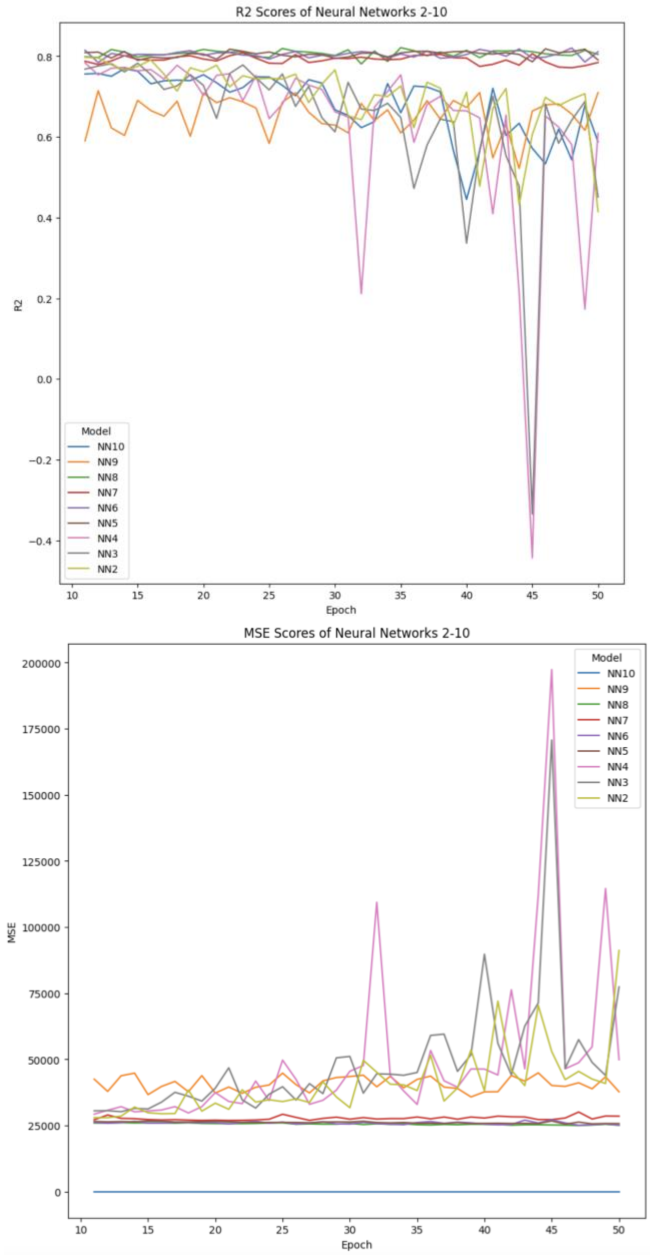


Figure 9: R^2 and MSE scores of Neural Network Architectures 2-10

There were some limitations to our methodology. The data we utilized was synthetically generated, meaning that there are likely more variations in actual real data. Hence our models would need to be retrained when we are able to capture that information. This could possibly result in slightly different outcomes. Additionally, because of time constraints, we were not able to explore more options within this space, as done by other researchers (Shapi et al., 2021).

However, we were able to answer our research question about whether synthetic data scaling for the purposes of electricity consumption works for training ML/AI models. We were also able to determine that the best model for our synthetic data when predicting electricity consumption was XGBoost given the US demographics.

In conclusion, to help Texas residents reduce their power costs and reduce the environmental impact of power generation and consumption, we sought to find a machine learning model that could predict the energy consumption of a customer. We tested five different types of models, Linear Regression, Support Vector Regression (SVR), XGBoost, Random Forest, and Neural Networks. These models were scored using the R^2 and the MSE loss metrics, and each model was trained twice using two different methodologies, fitting the model once and KFold validation using 10 folds. We found that, by using R^2 and MSE as metrics, XGBoost and Random Forest performed the best out of all of the models that were trained, with an R^2 score of 0.904 for XGBoost and an MSE score of 15,650.319 during 10 KFold Validation.

REFERENCES

[1] Trotta G. Electricity awareness and consumer demand for information. Int J Consum Stud. 2021; 45: 65–79. https://doi.org/10.1111/ijcs.12603

[2] Georgiou, A., Ioannou, P.A., & Christodoulides, P. (2013). Domestic Electricity Consumption and the Public Awareness Factor.

[3] Mulvad, L., Katan, S.H., Sundahl, N.V., & Jensen, L.I. (2012). Researching Motivational Factors Towards a Sustainable Electricity Consumption.

[5] Malik, M.A., & Kamarudin, M.F. (2020). Energy Meter Using a Smartphone.

[6] Bohdanowicz, Zbigniew, Beata Łopaciuk-Gonczaryk, Jarosław Kowalski, and Cezary Biele (2021). "Households’ Electrical Energy Conservation and Management: An Ecological Break-Through, or the Same Old Consumption-Growth Path?" Energies 14, no. 20: 6829. https://doi.org/10.3390/en14206829

[7] Fischer, C. (2008). Feedback on household electricity consumption: a tool for saving energy? Energy Efficiency, 1, 79-104.

[8] Jin, Y., Yang, E.J., & Fulton, J. (2021). An Empirical Study of Environmental Data Prediction in the United States Energy-Water Nexus. IEEE Access, 9, 32747-32759.

[9] Rashid, H., Mammen, P.M., Singh, S., Ramamritham, K., Singh, P., & Shenoy, P.J. (2017). Want to reduce energy consumption?: don't depend on the consumers! Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments.

[10] Santarossa, M., Das, N., Helwig, A., & Ahfock, T. (2016). Energy management and automated analytics for reduction of energy consumption. 2016 Australasian Universities Power Engineering Conference (AUPEC), 1-5.

[11] Coleman, M., Brown, N., Wright, A. J., & Firth, S. (2012). Information, communication and entertainment appliance use—Insights from a UK household study. Energy and Buildings, 54, 61–72. https://doi.org/10.1016/j.enbuild.2012.06.008

[12] Jones, J., Navvab, M., & Hill, Y. (1992). Operation and climate impact on electrical demand for institutional buildings. Conference Record of the 1992 IEEE Industry Applications Society Annual Meeting, 1852-1857 vol.2.

[13] Bennet, I.E., & O’brien, W. (2017). Office building plug and light loads: Comparison of a multi-tenant office tower to conventional assumptions. Energy and Buildings, 153, 461-475.

[14] Oar, Oap, & Cppd (2015). Reduce the Environmental Impact of Your Energy Use.

[15] Lehner, M., Mont, O., & Heiskanen, E. (2016). Nudging – A promising tool for sustainable consumption behaviour? Journal of Cleaner Production, 134, 166–177. https://doi.org/10.1016/j.jclepro.2015.11.086

[16] Laica, I., Blumberga, A., Rošā, M., & Blumberga, D. (2014). Determinants of household electricity consumption savings: A Latvian case study. Agronomy research, 12, 527-542.

[17] Fischer, C. Feedback on household electricity consumption: a tool for saving energy?. Energy Efficiency 1, 79–104 (2008). https://doi.org/10.1007/s12053-008-9009-7

[18] Chandrasekaran, S., Masthan, M., R, M.T., Athinarayanan, S., Ram, A., & V, V. (2023). Uncertainty-Aware Functional Analysis for Electricity Consumption Prediction Using Multi-Task Optimization Learning Model. 2023 International Conference on Computer Science and Emerging Technologies (CSET), 1-7.

[19] Wang, R., Lu, S., & Feng, W. (2020). A novel improved model for building energy consumption prediction based on model integration. Applied Energy, 262, 114561. https://doi.org/10.1016/j.apenergy.2020.114561

[20] Jovanović, R., Sretenović, A., & Živković, B. (2015). Ensemble of various neural networks for prediction of heating energy consumption. Energy and Buildings, 94, 189–199. https://doi.org/10.1016/j.enbuild.2015.02.052

[21] Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy and Buildings, 147, 77–89. https://doi.org/10.1016/j.enbuild.2017.04.038

[22] Wang, R., Lu, S., & Li, Q. (2019). Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings. Sustainable Cities and Society, 49, 101623. https://doi.org/10.1016/j.scs.2019.101623

[23] Østergård, T., Jensen, R. L., & Maagaard, S. (2018). A comparison of six metamodeling techniques applied to building performance simulations. Applied Energy, 211, 89–103. https://doi.org/10.1016/j.apenergy.2017.10.102

[24] Wei, Y., Zhang, X., Shi, Y., Liang, X., Pan, S., Wu, J., Han, M., & Zhao, X. (2018). A review of data-driven approaches for prediction and classification of building energy consumption. Renewable & Sustainable Energy Reviews, 82, 1027–1047. https://doi.org/10.1016/j.rser.2017.09.108

[25] Chai, T., & Draxler, R. R. (2014). "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature." Geoscientific Model Development, 7(3), 1247-1250.

[26] Glantz, S. A., & Slinker, B. K. (1990). "Primer of Applied Regression and Analysis of Variance." McGraw-Hill

[27] Pandi, Ganesh. (2018) USA Housing dataset: House Sales Price Predictions. https://www.kaggle.com/datasets/gpandi007/usa-housing-dataset?select=housing_test.csv

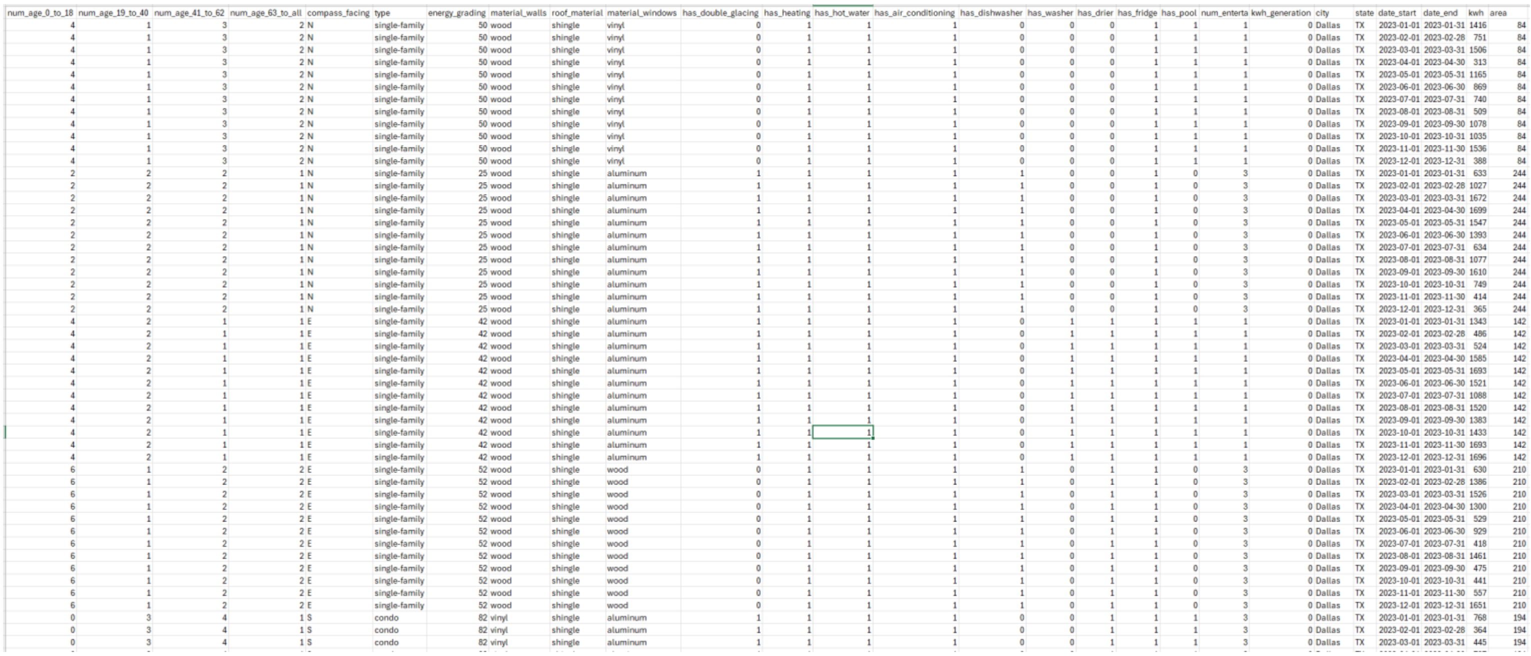
[28] Ang, Carmen. (2022) The Median Home Size in Every U.S. State in 2022. https://www.visualcapitalist.com/cp/median-home-size-every-american-state-2022/

[29] M. El Dayeh (2023) COVID-19_cases_plus_census.csv https://smu.instructure.com/courses/112577/files/8108114?wrap=1

[30] Douglass, M. (2020). Book Review: Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow, 2nd edition by Aurélien Géron. Physical and Engineering Sciences in Medicine/Physical and Engineering Sciences in Medicine, 43(3), 1135–1136. https://doi.org/10.1007/s13246-020-00913-z

[31] Shapi, M. K. M., Ramli, N. A., & Awalin, L. J. (2021). Energy consumption prediction by using machine learning for smart building: Case study in Malaysia. *Developments in the Built Environment*, *5*, 100037. https://doi.org/10.1016/j.dibe.2020.100037

Annex



num_age_0_to_18	num_age_19_to_40	num_age_41_to_62	num_age_63_to_all	compass_facing	type	energy_grading	material_walls	roof_material	material_windows	has_double_glacing	has_heating	has_hot_water	has_air_conditioning	has_dishwasher	has_washer	has_drier	has_fridge	has_pool	num_enterta	kwh_generation	city	state	date_start	date_end	kwh	area
4	1	3	2	N	single-family	50	wood	shingle	vinyl	0	1	1	1	0	0	0	1	1	1	0	Dallas	TX	2023-01-01	2023-01-31	1416	84
4	1	3	2	N	single-family	50	wood	shingle	vinyl	0	1	1	1	0	0	0	1	1	1	0	Dallas	TX	2023-02-01	2023-02-28	751	84
4	1	3	2	N	single-family	50	wood	shingle	vinyl	0	1	1	1	0	0	0	1	1	1	0	Dallas	TX	2023-03-01	2023-03-31	1506	84
4	1	3	2	N	single-family	50	wood	shingle	vinyl	0	1	1	1	0	0	0	1	1	1	0	Dallas	TX	2023-04-01	2023-04-30	313	84
4	1	3	2	N	single-family	50	wood	shingle	vinyl	0	1	1	1	0	0	0	1	1	1	0	Dallas	TX	2023-05-01	2023-05-31	1165	84
4	1	3	2	N	single-family	50	wood	shingle	vinyl	0	1	1	1	0	0	0	1	1	1	0	Dallas	TX	2023-06-01	2023-06-30	869	84
4	1	3	2	N	single-family	50	wood	shingle	vinyl	0	1	1	1	0	0	0	1	1	1	0	Dallas	TX	2023-07-01	2023-07-31	740	84
4	1	3	2	N	single-family	50	wood	shingle	vinyl	0	1	1	1	0	0	0	1	1	1	0	Dallas	TX	2023-08-01	2023-08-31	509	84
4	1	3	2	N	single-family	50	wood	shingle	vinyl	0	1	1	1	0	0	0	1	1	1	0	Dallas	TX	2023-09-01	2023-09-30	1078	84
4	1	3	2	N	single-family	50	wood	shingle	vinyl	0	1	1	1	0	0	0	1	1	1	0	Dallas	TX	2023-10-01	2023-10-31	1035	84
4	1	3	2	N	single-family	50	wood	shingle	vinyl	0	1	1	1	0	0	0	1	1	1	0	Dallas	TX	2023-11-01	2023-11-30	1536	84
4	1	3	2	N	single-family	50	wood	shingle	vinyl	0	1	1	1	0	0	0	1	1	1	0	Dallas	TX	2023-12-01	2023-12-31	388	84
2	2	2	1	N	single-family	25	wood	shingle	aluminum	1	1	1	1	1	0	0	1	0	3	0	Dallas	TX	2023-01-01	2023-01-31	633	244
2	2	2	1	N	single-family	25	wood	shingle	aluminum	1	1	1	1	1	0	0	1	0	3	0	Dallas	TX	2023-02-01	2023-02-28	1027	244
2	2	2	1	N	single-family	25	wood	shingle	aluminum	1	1	1	1	1	0	0	1	0	3	0	Dallas	TX	2023-03-01	2023-03-31	1672	244
2	2	2	1	N	single-family	25	wood	shingle	aluminum	1	1	1	1	1	0	0	1	0	3	0	Dallas	TX	2023-04-01	2023-04-30	1699	244
2	2	2	1	N	single-family	25	wood	shingle	aluminum	1	1	1	1	1	0	0	1	0	3	0	Dallas	TX	2023-05-01	2023-05-31	1547	244
2	2	2	1	N	single-family	25	wood	shingle	aluminum	1	1	1	1	1	0	0	1	0	3	0	Dallas	TX	2023-06-01	2023-06-30	1393	244
2	2	2	1	N	single-family	25	wood	shingle	aluminum	1	1	1	1	1	0	0	1	0	3	0	Dallas	TX	2023-07-01	2023-07-31	634	244
2	2	2	1	N	single-family	25	wood	shingle	aluminum	1	1	1	1	1	0	0	1	0	3	0	Dallas	TX	2023-08-01	2023-08-31	1077	244
2	2	2	1	N	single-family	25	wood	shingle	aluminum	1	1	1	1	1	0	0	1	0	3	0	Dallas	TX	2023-09-01	2023-09-30	1610	244
2	2	2	1	N	single-family	25	wood	shingle	aluminum	1	1	1	1	1	0	0	1	0	3	0	Dallas	TX	2023-10-01	2023-10-31	749	244
2	2	2	1	N	single-family	25	wood	shingle	aluminum	1	1	1	1	1	0	0	1	0	3	0	Dallas	TX	2023-11-01	2023-11-30	414	244
2	2	2	1	N	single-family	25	wood	shingle	aluminum	1	1	1	1	1	0	0	1	0	3	0	Dallas	TX	2023-12-01	2023-12-31	365	244
4	2	1	1	E	single-family	42	wood	shingle	aluminum	1	1	1	1	0	1	1	1	1	1	0	Dallas	TX	2023-01-01	2023-01-31	1343	142
4	2	1	1	E	single-family	42	wood	shingle	aluminum	1	1	1	1	0	1	1	1	1	1	0	Dallas	TX	2023-02-01	2023-02-28	486	142
4	2	1	1	E	single-family	42	wood	shingle	aluminum	1	1	1	1	0	1	1	1	1	1	0	Dallas	TX	2023-03-01	2023-03-31	524	142
4	2	1	1	E	single-family	42	wood	shingle	aluminum	1	1	1	1	0	1	1	1	1	1	0	Dallas	TX	2023-04-01	2023-04-30	1585	142
4	2	1	1	E	single-family	42	wood	shingle	aluminum	1	1	1	1	0	1	1	1	1	1	0	Dallas	TX	2023-05-01	2023-05-31	1693	142
4	2	1	1	E	single-family	42	wood	shingle	aluminum	1	1	1	1	0	1	1	1	1	1	0	Dallas	TX	2023-06-01	2023-06-30	1521	142
4	2	1	1	E	single-family	42	wood	shingle	aluminum	1	1	1	1	0	1	1	1	1	1	0	Dallas	TX	2023-07-01	2023-07-31	1088	142
4	2	1	1	E	single-family	42	wood	shingle	aluminum	1	1	1	1	0	1	1	1	1	1	0	Dallas	TX	2023-08-01	2023-08-31	1520	142
4	2	1	1	E	single-family	42	wood	shingle	aluminum	1	1	1	1	0	1	1	1	1	1	0	Dallas	TX	2023-09-01	2023-09-30	1383	142
4	2	1	1	E	single-family	42	wood	shingle	aluminum	1	1	1	1	0	1	1	1	1	1	0	Dallas	TX	2023-10-01	2023-10-31	1433	142
4	2	1	1	E	single-family	42	wood	shingle	aluminum	1	1	1	1	0	1	1	1	1	1	0	Dallas	TX	2023-11-01	2023-11-30	1693	142
4	2	1	1	E	single-family	42	wood	shingle	aluminum	1	1	1	1	0	1	1	1	1	1	0	Dallas	TX	2023-12-01	2023-12-31	1696	142
6	1	2	2	E	single-family	52	wood	shingle	wood	0	1	1	1	1	0	1	1	0	3	0	Dallas	TX	2023-01-01	2023-01-31	630	210
6	1	2	2	E	single-family	52	wood	shingle	wood	0	1	1	1	1	0	1	1	0	3	0	Dallas	TX	2023-02-01	2023-02-28	1386	210
6	1	2	2	E	single-family	52	wood	shingle	wood	0	1	1	1	1	0	1	1	0	3	0	Dallas	TX	2023-03-01	2023-03-31	1526	210
6	1	2	2	E	single-family	52	wood	shingle	wood	0	1	1	1	1	0	1	1	0	3	0	Dallas	TX	2023-04-01	2023-04-30	1300	210
6	1	2	2	E	single-family	52	wood	shingle	wood	0	1	1	1	1	0	1	1	0	3	0	Dallas	TX	2023-05-01	2023-05-31	529	210
6	1	2	2	E	single-family	52	wood	shingle	wood	0	1	1	1	1	0	1	1	0	3	0	Dallas	TX	2023-06-01	2023-06-30	929	210
6	1	2	2	E	single-family	52	wood	shingle	wood	0	1	1	1	1	0	1	1	0	3	0	Dallas	TX	2023-07-01	2023-07-31	418	210
6	1	2	2	E	single-family	52	wood	shingle	wood	0	1	1	1	1	0	1	1	0	3	0	Dallas	TX	2023-08-01	2023-08-31	1461	210
6	1	2	2	E	single-family	52	wood	shingle	wood	0	1	1	1	1	0	1	1	0	3	0	Dallas	TX	2023-09-01	2023-09-30	475	210
6	1	2	2	E	single-family	52	wood	shingle	wood	0	1	1	1	1	0	1	1	0	3	0	Dallas	TX	2023-10-01	2023-10-31	441	210
6	1	2	2	E	single-family	52	wood	shingle	wood	0	1	1	1	1	0	1	1	0	3	0	Dallas	TX	2023-11-01	2023-11-30	557	210
6	1	2	2	E	single-family	52	wood	shingle	wood	0	1	1	1	1	0	1	1	0	3	0	Dallas	TX	2023-12-01	2023-12-31	1651	210
0	3	4	1	S	condo	82	vinyl	shingle	aluminum	1	1	1	1	0	0	1	1	1	3	0	Dallas	TX	2023-01-01	2023-01-31	768	194
0	3	4	1	S	condo	82	vinyl	shingle	aluminum	1	1	1	1	0	0	1	1	1	3	0	Dallas	TX	2023-02-01	2023-02-28	364	194
0	3	4	1	S	condo	82	vinyl	shingle	aluminum	1	1	1	1	0	0	1	1	1	3	0	Dallas	TX	2023-03-01	2023-03-31	445	194