California Energy Consumption Prediction

Frank Dong, 91576270, zhuojiad@uci.edu Sean Lee, 66676792, seanyl5@uci.edu
Nima Hendi, 14209313, nhendi@uci.edu Yang Weng, 61484846,
yweng10@uci.edu

May 24, 2023

Abstract

The project aims to predict energy consumption in California based on various price data using different machine learning models. Specifically, SARIMAX, Linear, Dense, Multi-Step Dense, and LSTM models were developed and evaluated for their predictive performance on energy consumption data. Predictions were made for 2 years, 5 years, and 10 years into the future, taking into account factors such as trends and seasonality, price data, policy changes, and technological advancements. The results of the project can be useful for energy companies, policymakers, and other stakeholders in planning for future energy demand and supply, and in developing more efficient and sustainable energy systems.

1 Introduction and Problem Statement

Energy consumption is a key indicator of economic activity and societal development, and accurate forecasting of energy demand is essential for effective energy planning and manage ment. In this project, we focus on predicting energy consumption in California, the largest energy-consuming state in the United States, using various price data and machine learning models.

There are great amount of data available from different reliable sources such as U.S. En ergy Information Administration (eia.gov) and California's official government website (en ergy.ca.gov). This data encompasses information on electricity usage, natural gas consump tion, and renewable energy production over time. By analyzing historical consumption pat terns, we can identify seasonal variations, peak demand periods, and long-term trends, which are essential for accurate predictions. Furthermore, We used and analysed various auxiliary data sources contribute to energy consumption predictions in California. These include de mographic data, such as population density, housing types, household sizes and weather and climate data.

We develop and evaluate several machine learning models, including SARIMAX, Linear, Dense, Multi-Step Dense, and LSTM models, using appropriate validation and testing meth ods. By comparing the performance of these models, we aim to identify the most effective approaches for predicting energy consumption in California over the next 2, 5, and 10 years.

The insights gained from this project can be useful for energy companies,

policymakers, and 1

other stakeholders in planning for future energy demand and supply, and in developing more efficient and sustainable energy systems. Additionally, the project contributes to the growing body of research on time series forecasting and machine learning, with potential applications in other fields beyond energy consumption prediction.

2 Related Work

The field of predictive analysis has seen significant growth over the past few years, with researchers and data scientists seeking to develop models that can accurately forecast future trends and patterns. Over the years, researchers have employed various methodologies to accurately predict energy consumption in different regions.

In the case of Zheng-Xin Wang's article, ”A Predictive Analysis of Clean Energy Consump tion, Economic Growth and Environmental Regulation in China Using an Optimized Grey Dynamic Model,” a grey dynamic model, statistical and computational modeling that com bines a partial theoretical structure with related data to complete the model, was used to predict clean energy usage in China by the year 2020. This model is particularly useful for short-term problems with small samples and limited information, as it can effectively iden tify trends and patterns in the data. The core principle of the grey dynamic model lies in the concept of ”whitening” and ”blackening” the original data sequence. Whitening refers to the process of extracting useful information from the original data, while blackening involves the forecasting of future trends based on the extracted information.

In addition to the grey dynamic model, another notable approach in the field of predictive analysis is the CNN-LSTM model. The article ”Predicting residential energy consumption using CNN-LSTM neural networks” published in 2019 by Cho and Kim utilizes the CNN LSTM model to predict residential energy consumption. It is a hybrid neural network ar chitecture that combines the strengths of both CNNs and LSTMs. This model is commonly used for analyzing and processing sequential data, particularly in tasks that involve capturing spatial and temporal dependencies, such as image or time series analysis. The convolutional layers of the CNN component are employed to process the input data and extract meaningful spatial features. The LSTM component of the model utilizes memory cells and gates to han dle the sequential nature of the data. These gates regulate the flow of information, enabling the LSTM to selectively retain or discard information at each time step.

In conclusion, the field of predictive analysis in energy consumption has seen significant growth, with the grey dynamic model offering a valuable solution for short-term forecasting with limited data, and the CNN-LSTM model providing an effective approach for capturing spatial and temporal dependencies in sequential

data. These methodologies showcase the diverse range of techniques employed in predicting energy consumption patterns, and as re searchers continue to refine these approaches, further advancements in accurate forecasting are expected.

3 Data Sets

In this project, we will be using data from the U.S. Energy Information Administration web site (https://www.eia.gov/totalenergy/data) and the California Energy Commission website

as (https://www.energy.ca.gov/data-reports). We will be working on datasets that provide monthly energy usage, effects of temperature, and different needs of resources for different sectors. The data sets will be in CSV and EXCEL formats.

In this project, we employ feature selection techniques to identify the most important features for predicting energy consumption in California. To achieve this, we first calculate the cor relation between each feature and the target variable (energy consumption). We select only those features with a correlation coefficient greater than 0.85, as these are considered highly correlated with the target variable.

Additionally, we eliminate duplicate features that are highly correlated with each other, as these can lead to multicollinearity issues and reduce the stability of the predictive models. Specifically, we remove one of TNASB and PEASB, as well as OPSCB and OPISB, as these pairs of features are highly correlated with each other.

After applying these feature selection techniques, we are able to reduce the number of fea tures from 306 down to 11, which can help to improve the efficiency and accuracy of the predictive models. By focusing on the most important features, we can better understand the factors that influence energy consumption in California and develop more effective strategies for managing and planning for future energy demand.

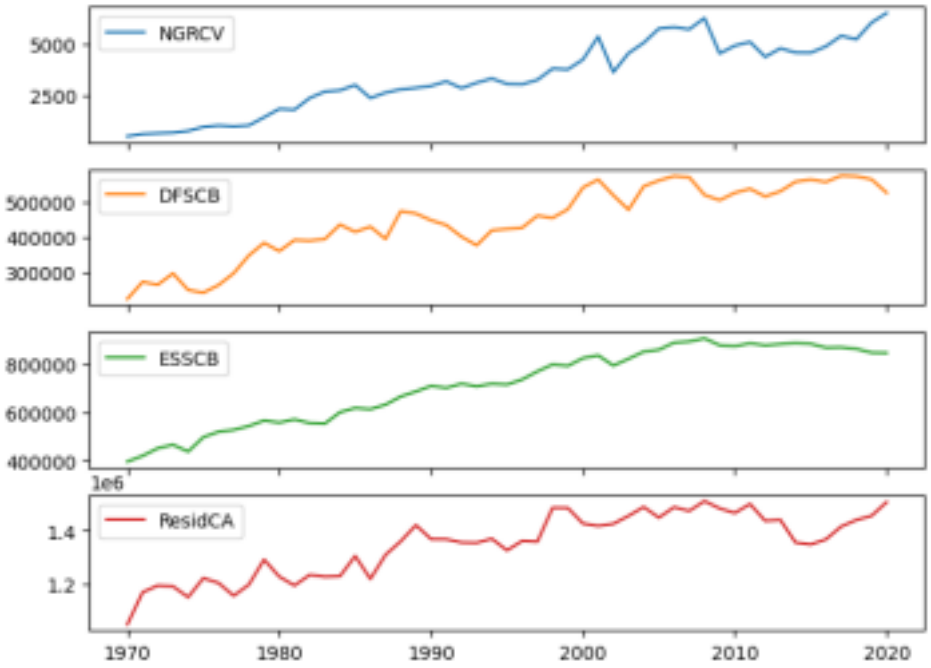


Figure 1: Three response variables vs predictor variable from 1970 to 2020

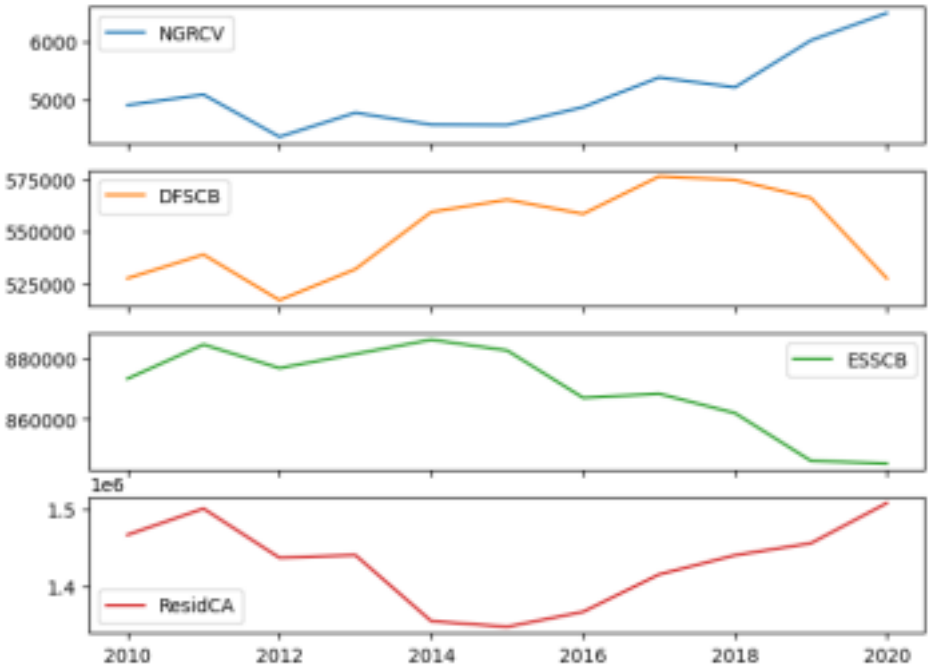


Figure 2: Three response variables vs predictor variable from 2010 to 2020

”DFSCB”:”Distillate fuel oil total consumption adjusted for process

fuel.” ”ESSCB”:”Electricity total consumption adjusted for process

fuel.”,

”NGRCV”:”Natural gas expenditures in the residential sector”,

”OPISB”:”Other petroleum products consumed by the industrial sector excluding refinery fuel.”,

”TEPFB”:”Total energy used as process fuel and other consumption that has no direct fuel costs”,

”TNSCB”:”Total net energy consumption”,

”WWIXB” : ”Wood and waste consumed by the industrial sector, at no cost.”

”PEASB” : ”Primary energy consumed by the transportation sector, adjusted for process fuel.”

4 Overall Technical Approach

For this project, First we split the data into three sets: training, validation, and test sets. We use a (70, 20, 10) percent split, where 70 percent of the data is used for training, 20 percent for validation, and 10 percent for testing.

It’s important to note that we do not randomly shuffle the data before splitting it. This is for two reasons. Firstly, it ensures that we can still chop the data into windows of consecutive samples, which is important for time-series analysis. Secondly, it ensures that the validation

and test results are more realistic, as they are evaluated on data collected after the model was trained.

By using this split, we can train our models on a large portion of the data while still reserving a significant portion for validation and testing. This allows us to accurately evaluate the performance of our models and make more informed decisions about which models to use for predicting energy consumption in California.

Then, we apply normalization to the data before training our neural network models. Scaling the features in this way is important because it helps to ensure that each feature contributes equally to the model and prevents any particular feature from dominating the learning pro cess.

Normalization is carried out by subtracting the mean and dividing by the standard deviation of each feature. It’s important to note that the mean and standard deviation should only be computed using the training data. This ensures that the models have no access to the values in the validation and test sets, which helps to

prevent overfitting and improves the generalization performance of the models.

By applying normalization to the data, we can improve the efficiency and accuracy of our neural network models, making it easier to identify the most important features for predicting energy consumption in California.

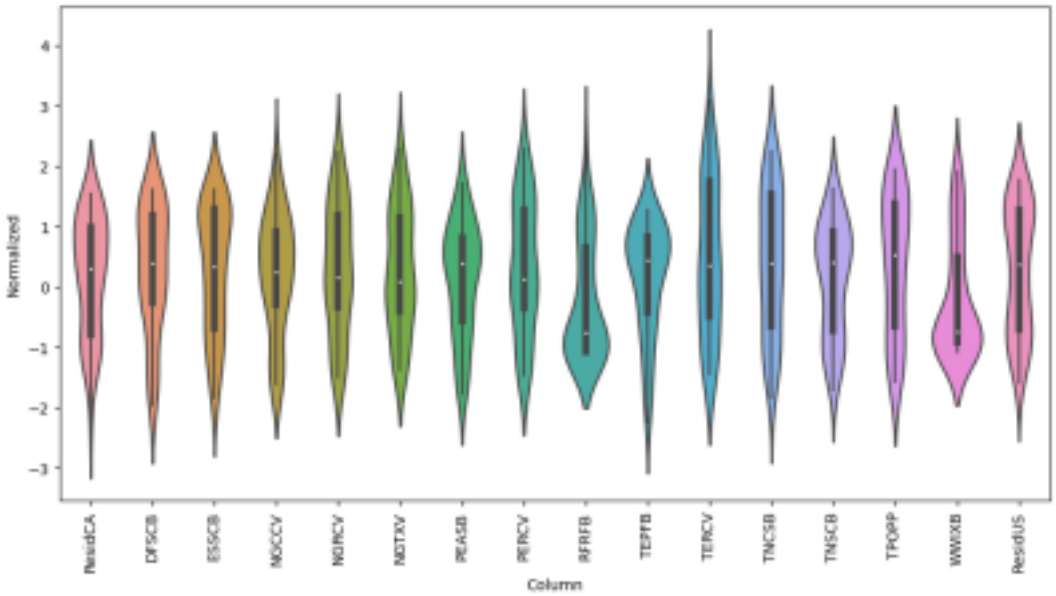


Figure 3: Normalization

we use a windowing technique to prepare the data for training our models. This involves creating windows of consecutive samples from the data, which are used as input and output sequences for our neural network models.

The key features of the input windows include the width, or number of time steps, of the input and label windows, as well as the time offset between them. Additionally, we must specify which features are used as inputs, labels, or both, which helps to ensure that the models are trained on the most relevant and important data.

By using data windowing, we can better understand how the various features of the data contribute to energy consumption in California over time, and develop more effective models for predicting future energy demand.

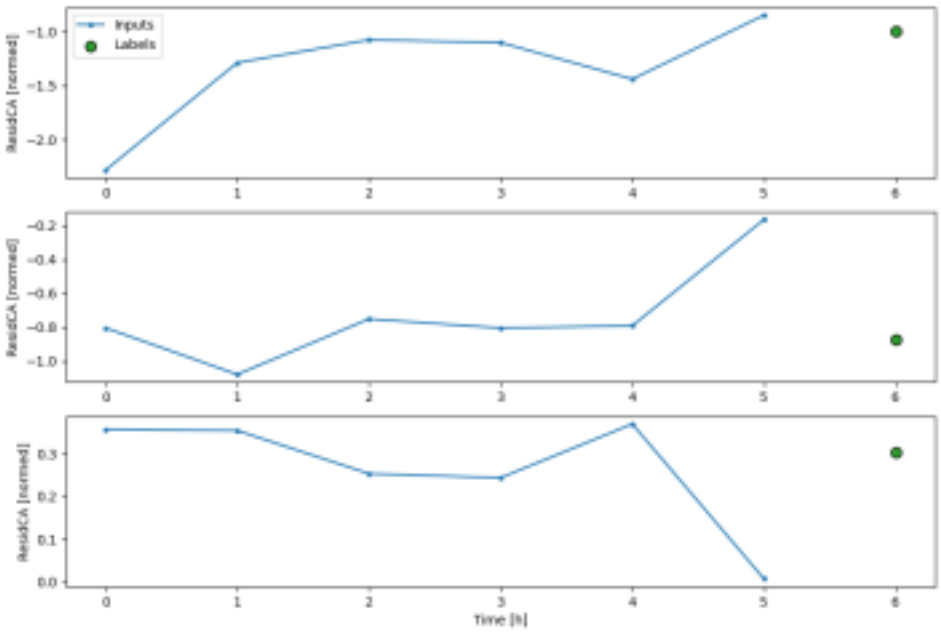


Figure 4: Three different input data sample

In addition to the linear model, we can also apply a variety of other machine learning models to the task of predicting energy consumption in California. These include the dense neural network model, the long short-term memory (LSTM) model, and the seasonal autoregressive integrated moving average with exogenous variables (SARIMAX) model.

The dense neural network model is a type of feedforward neural network that can be used to model complex relationships between input features and output. By stacking multiple layers of neurons, the dense neural network is able to learn increasingly abstract representations of the input data, which can help to improve the accuracy of its predictions.

The LSTM model is a type of recurrent neural network that is designed to handle sequential data. It is particularly useful for time-series analysis, as it can capture temporal dependencies between successive inputs and outputs. The LSTM model achieves this by incorporating a memory cell and various gating mechanisms that allow it to selectively forget or remember information from previous time steps.

The SARIMAX model is a classical statistical model that is widely used for time-series fore casting. It is a variant of the ARIMA model that includes additional exogenous variables, which can help to capture the effects of external factors on the time series. By fitting a SARIMAX model to the energy consumption data, we can gain insights into the underlying patterns and trends that are driving energy consumption in California.

By comparing the performance of these different models on the task of predicting energy consumption, we can gain a better understanding of the strengths and limitations of different

approaches to time-series analysis and modeling. This can help us to identify the most effec tive techniques for predicting energy demand in California, and develop more accurate and reliable forecasts for the future.

5 Software

Our group utilized Jupyter Notebook to code and GitHub to collaborate. Jupyter Notebook provides a user-friendly interface that allows our team members to work on different parts of the project simultaneously. By breaking up the code into different sections, our team was able to easily select and run only the codes that need to be debugged or fixed, which saves time and improves productivity. Furthermore, the ability to run the models several times without having to wait for the entire file to run is a huge advantage in the data analysis process.

Once the data has been cleaned and linked together by their year and MSN index, it is passed into PostgreSQL for storage. The pandas library in Python is then used to read the data back into Jupyter Notebook, where it can be manipulated and analyzed. The data is then fed into the ARIMA and SARIMAX models, which were imported from the statsmodels library. The statsmodels library is a powerful tool for data analysis, as it contains a wide range of classes and functions for numerous estimations and statistical models. Although the main model used in the project is the SARIMAX model, ARIMA was also imported since SARIMAX is a variation of the ARIMA model.

In addition to the matplotlib library, our team also utilized the seaborn library for graphics. Seaborn is a data visualization library based on matplotlib, but it provides a higher level of statistical graphics. Our team used this library to show the heat maps of correlation matrices, which are useful for identifying relationships between different variables in the data.

6 Experiments and Evaluation

SARIMAX
SARIMAX, short for Seasonal Autoregressive Integrated Moving Average with Exogenous Variables, is a time series forecasting model that extends the traditional ARIMA model to incorporate seasonality and exogenous variables. It is a powerful tool for analyzing and fore casting time series data that exhibits seasonal patterns and is influenced by external factors.

SARIMAX is generally preferred when the data exhibits clear seasonal patterns and there are known external factors that influence the time series.

It relies on the assumption that future values depend on past values and can capture linear relationships.

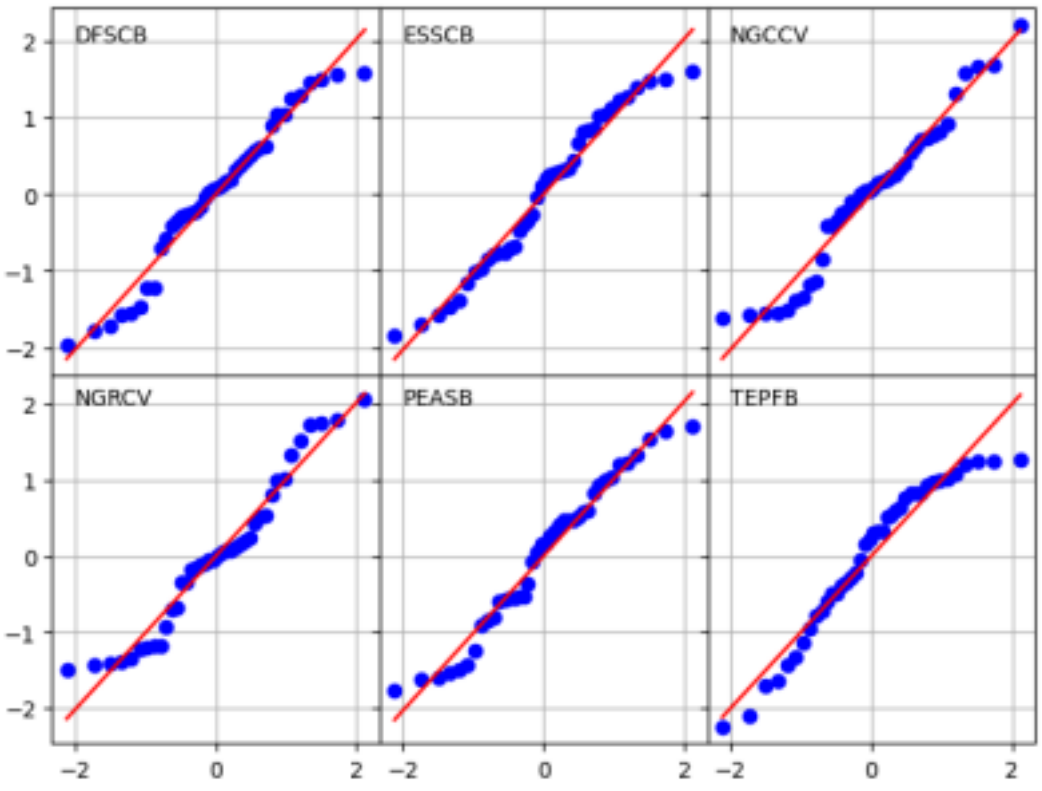


Figure 5: Linear relationship assumption

SARIMAX Prediction

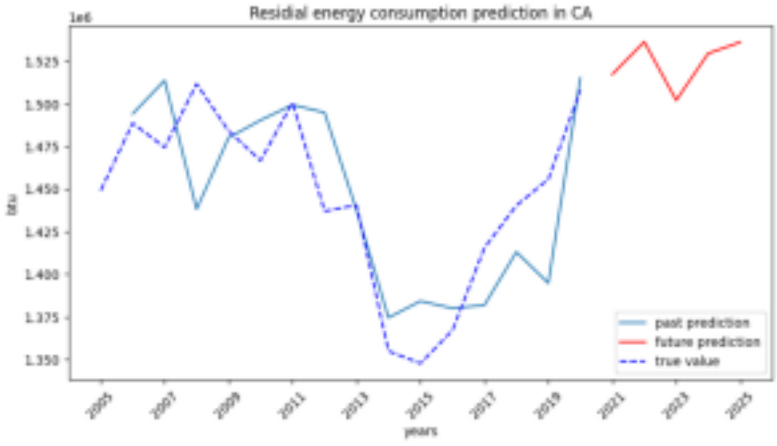


Figure 6: 5-year prediction with validation with validation error rate 1% 8

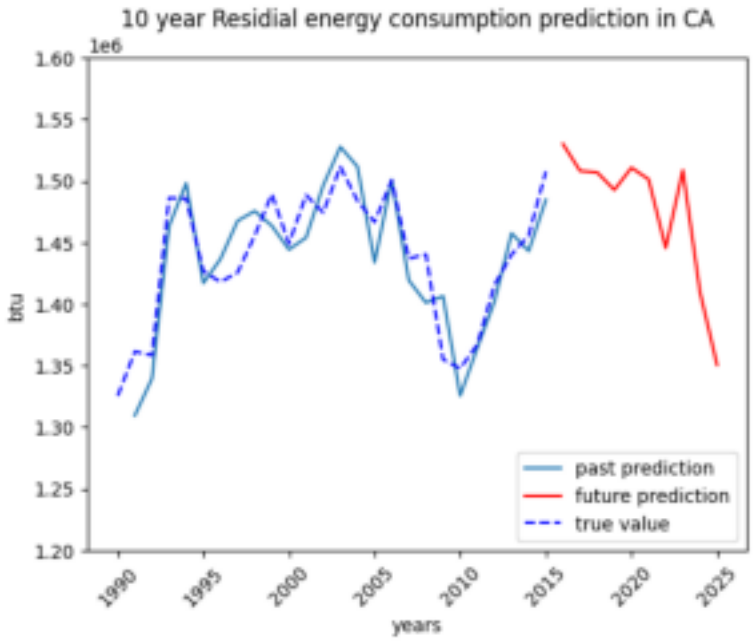


Figure 7: 10-year prediction with validation with validation error rate 3%

7 Notebook Description

In this notebook, we present an introduction to the data and models used for forecasting California's energy consumption. Our chosen model, SARIMAX, demonstrates accurate predictions, revealing a significant increase in residential energy consumption over the past five years. However, our forecast indicates a sharp decline in residential energy usage over the next decade, returning to levels similar to those in 2010. This underscores the need for sustainable energy practices and efficient solutions to align with the projected decrease in residential energy consumption, enabling stakeholders to proactively address future energy demands.

8 Members Participation

Frank Dong:
Data gathering, Data processing, SARIMAX and LSTM model training, validation, and vi sualization, and report structure assign.

Sean Lee:

Decision Tree model training, validation, and visualization. Report formatting.

Yang Weng:
Linear Regression model training, validation, and visualization. Report formatting. 9

9 Discussion and Conclusion

Key Outcomes

The analysis conducted in this study yielded several key outcomes. Firstly, it successfully identified an accurate model for predicting yearly data with a seasonal pattern, allowing for reliable projections in the future. The selected model, SARIMAX, outperformed other statis tical and machine learning models, demonstrating its effectiveness in capturing the underly ing patterns in the data.

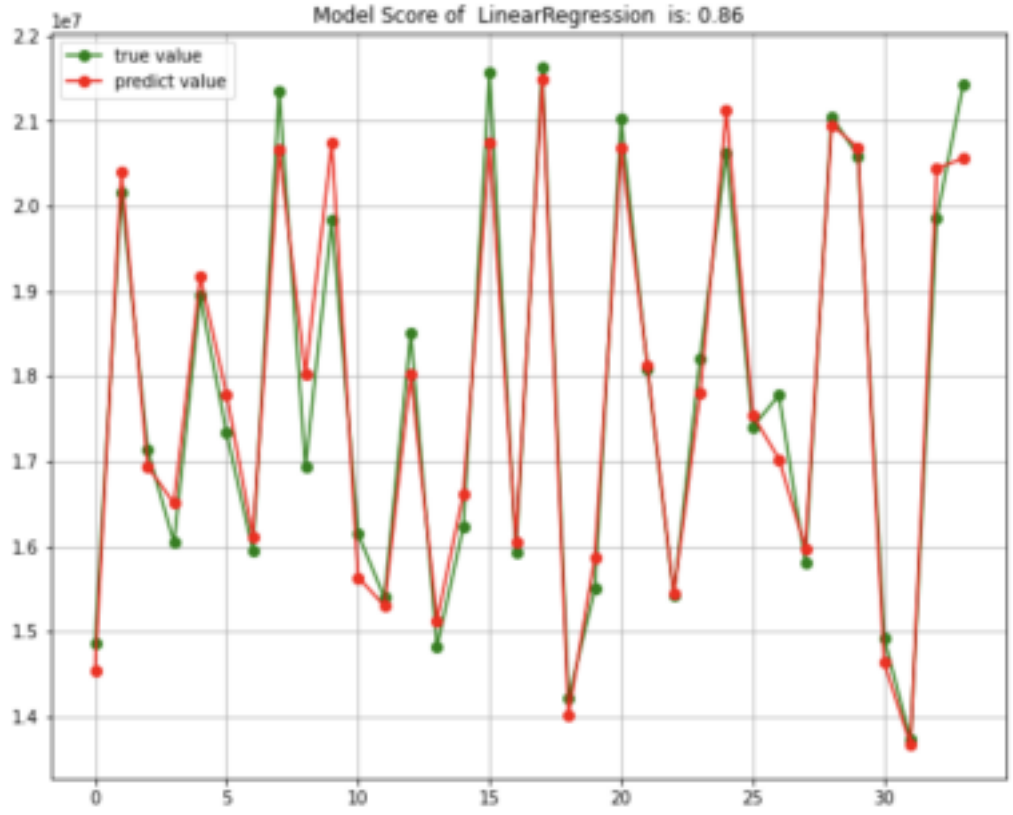
Furthermore, the applicability of this model extends beyond the specific energy sector and geographic region analyzed in this study. It can be applied to other energy sources, sectors, and even to other states within the United States. This versatility enhances the usefulness and potential impact of the model, allowing for comprehensive energy consumption forecasts in various contexts.

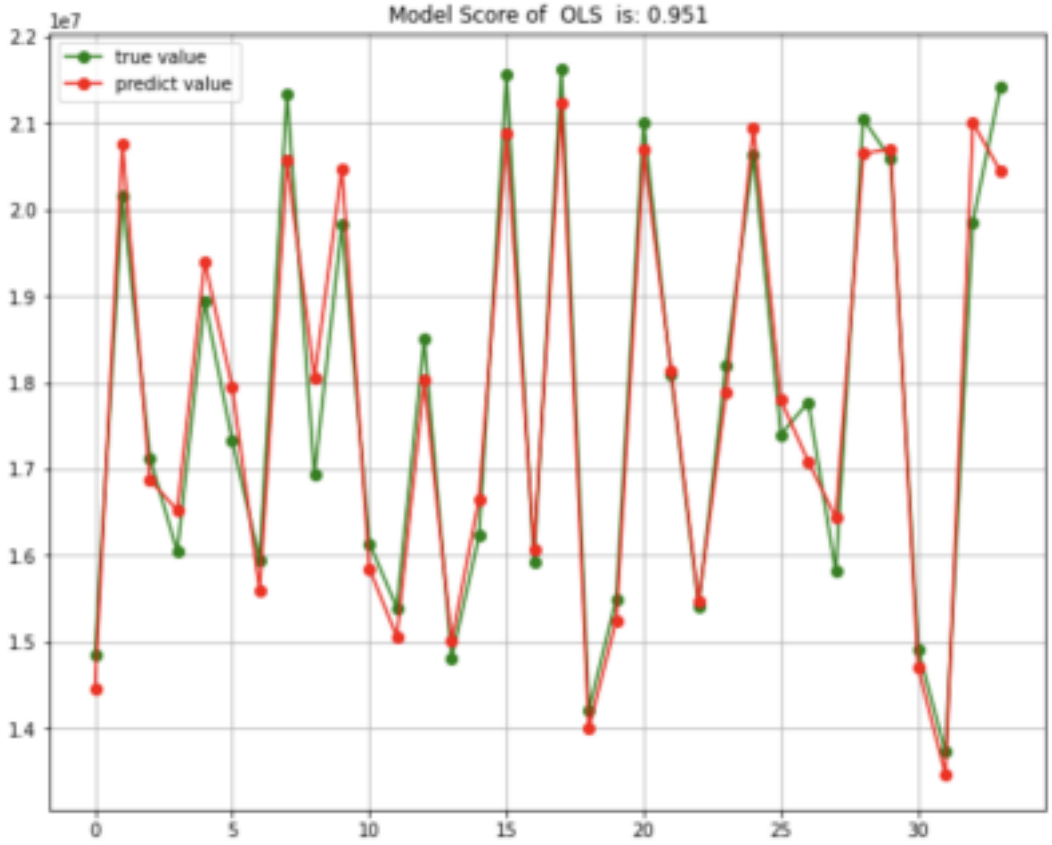
One significant finding of the analysis is that residential energy consumption remains consis tently high over a span of five years. However, the study also indicates a sharp decrease in residential energy consumption is expected in the next ten years. This projection suggests po tential shifts in energy usage patterns and highlights the importance of developing sustainable and efficient energy solutions to meet future demands.

Among the identified factors influencing energy consumption, six features emerged as the most influential. These factors include resident population, electric total consumption, natural gas expenditures, total net energy consumption, natural gas expenditures in the residential sector, and natural gas expenditures in the commercial sector. This implies that natural gas continues to be a crucial energy source in the daily lives of California residents, underscoring the need for effective energy management strategies and policies.

In summary, this analysis provides a robust model for yearly data prediction with a seasonal pattern and identifies key factors driving energy consumption. The findings are not only applicable to the analyzed energy sector and geographic region but can also be extended to other sectors, energy sources, and states. The insights gained from this study contribute to a better understanding of energy usage trends and facilitate informed decision-making in the field of energy management and policy.

Appendix


Model Score of LinearRegression is: 0.86
1e7
true value
predict value
2.2
2.1
2.0
1.9
1.8
1.7
1.6
1.5
1.4
0
5
10
15
20
25
30



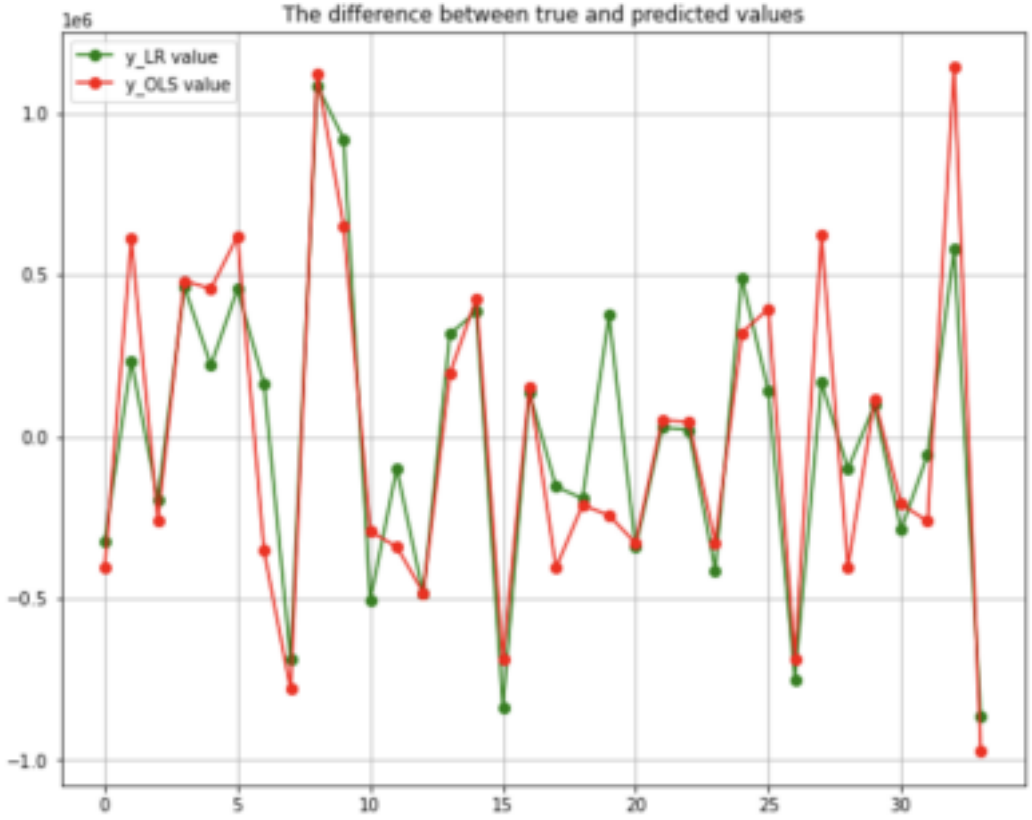


Figure 8:

We aimed to develop a regression model to predict the target variable 'ResidUS' using a given dataset. The dataset contained multiple features, and we aimed to identify the most relevant features for predicting the target variable. We followed a

systematic approach to achieve our goal, which included preprocessing steps, data management, feature selection, and model development. We also compared two different regression models to determine the best one for our prediction task.

We started by loading the dataset and creating a correlation matrix to identify the features that are highly correlated with the target variable 'ResidCA'. We used a correlation threshold of 0.85 to select the most relevant features. After identifying the highly correlated features, we removed 'ResidCA' from the list of high correlation features and selected the remaining high correlation features for further analysis.

We removed 'ResidUS' from the list of high correlation features as it is our target variable. We then created a new DataFrame called 'df high corr', which contains only the target vari able 'ResidUS' and the selected high correlation features. This DataFrame was used to train our regression models.

For the development of the regression model, we employed two methods: Linear Regression and Ordinary Least Squares (OLS). We split our dataset into training and testing sets using a 70/30 ratio. The training set was used to fit the models, and the test set was used to evaluate their performance.

We employed the Linear Regression model from the 'sklearn.linear model' library. We fitted the model on the training set and predicted the target variable 'ResidUS' on the test set. We calculated the model's performance metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared. We plotted the true values against the predicted values to visually inspect the model's performance.

We used the OLS model from the 'statsmodels.api' library as an alternative regression model. The process of fitting the model and predicting the target variable 'ResidUS' was similar to the Linear Regression method. We also computed the performance metrics and plotted the true values against the predicted values.

To compare the performance of the two regression models, we calculated the difference be tween the true and predicted values for each model. We plotted these differences to visually inspect which model performed better in predicting the target variable 'ResidUS'.

In conclusion, we followed a structured approach to preprocess the data, manage the features, develop regression models, and compare their performance. This allowed us to identify the best regression model for predicting the target variable 'ResidUS' based on our dataset.