# California Energy Consumption Prediction

Frank Dong, 91576270, zhuojiad@uci.edu
Sean Lee, 66676792, seanyl5@uci.edu

April 27, 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 management. 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.

The dataset contains a large number of features but only 60 time points, which presents a challenge for developing accurate predictive models. We address this challenge by employing feature selection techniques and incorporating external data sources, such as historical energy consumption and weather patterns, to supplement the limited time series data.

We develop and evaluate several machine learning models, including SARIMAX, Linear, Dense, Multi-Step Dense, and LSTM models, using appropriate validation and testing methods. 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 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. In the case of Zheng-Xin Wang's article, "A Predictive Analysis of Clean Energy Consumption, Economic Growth and Environmental Regulation in China Using an Optimized Grey Dynamic Model," a grey dynamic 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 identify trends and patterns in the data.

However, while the grey dynamic model may have been effective for Wang's analysis, it may not be the ideal model for every dataset or problem. In fact, when it comes to predicting long-term trends or patterns, the grey dynamic model may not be the most appropriate choice. This is because such models rely heavily on historical data and may not be able to accurately account for changes or developments that may occur in the future. In addition, when analyzing larger datasets, it may be more beneficial to use other types of models, such as machine learning algorithms, which can identify complex patterns and relationships in the data.

Despite its limitations, the grey dynamic model remains a valuable tool for many researchers and data scientists, particularly those working with smaller datasets or short-term problems. By utilizing this model, they can make accurate predictions based on limited information and identify trends that may not be visible through other means. Ultimately, the choice of model will depend on the specific needs and goals of the analysis, as well as the nature of the data being analyzed.

## 3 Data Sets

In this project, we will be using data from the U.S. Energy Information Administration website (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 correlation 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 features 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.

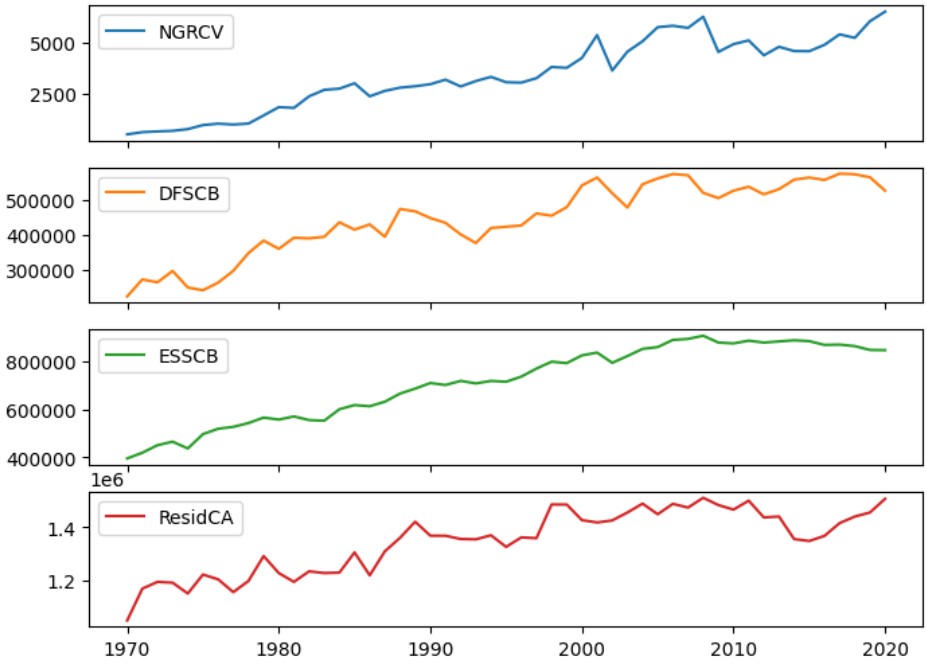


Figure 1:

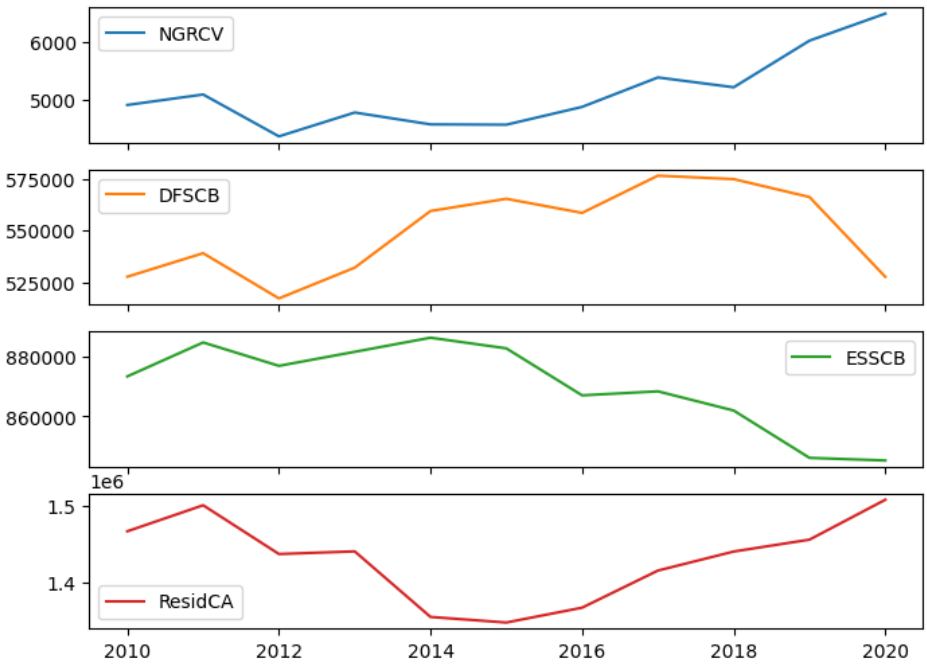


Figure 2:

”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 process.

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.

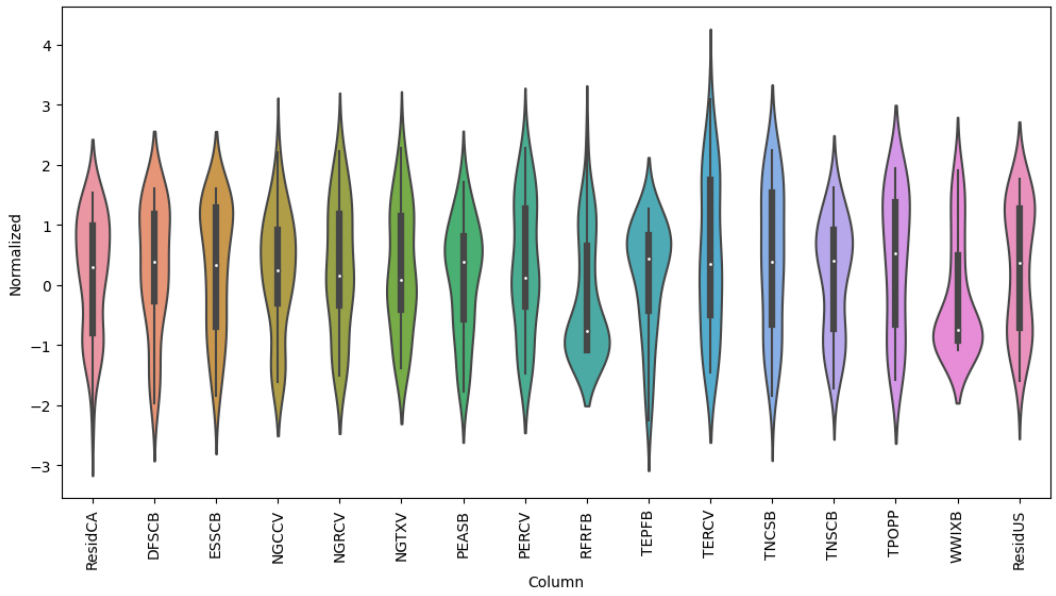


Figure 3: Normalization

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.

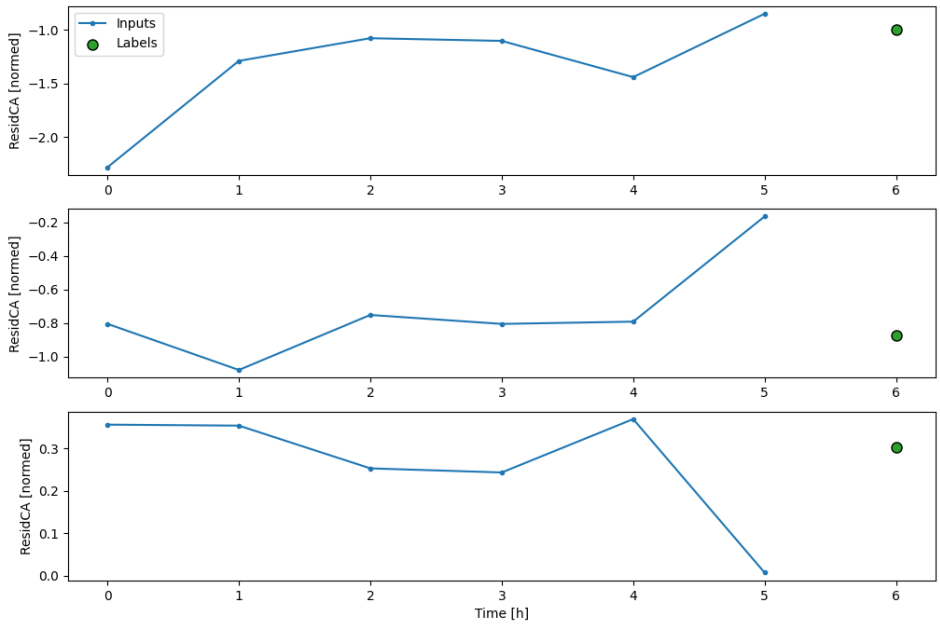


Figure 4: Three different input data sample

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 forecasting. 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 effective 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

# 7 Notebook Description

# 8 Members Participation

# 9 Discussion and Conclusion