



Predicting Energy Usage in California

-UCI Capstone 2023

Frank Dong, Yang Weng, Nima Hendi, Sean Lee



Introduction

- Project objective: clear statement of the problem
- Available data + sources

In this project, we want to develop a system to forecast energy usage per sector (residential, commercial, industrial) and per source for California.




Project Summary (California)

Based on historical data from 1960 to 2020, from CA and US. (more than 10,000 features)

Prediction for the next 2 year (2020-2022), 5 year (2020-2025), and 10 year (2020-2030).

Prediction for energy uses for Residential, commercial, industrial, Transportation and Total consumption.

Prediction for energy uses for Coal, Natural Gas, Petroleum, Nuclear, Renewable Energy.



Data Description and Management

Link to data slide decks: [data uci capstone](#)

Predicting energy uses:

<https://docs.google.com/presentation/d/1zvVwxJ04zf214ujSyKs-WahalehKlI8IFVgxumDou5M/edit?usp=sharing>

[SQL database Github Repository](#)

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



Database (PostgreSQL)

- Database store more than ten thousands different features, combined from different category of energy uses, and price.
- Data are linked together by MSN index, MSN translation included in data file.
- Independent variable: price and consumption table named (All and Price_CA)
- Target variable: 5 energy source table and 5 energy sector table.



Data Pipeline

- Concatenate target variable and response variable
- Transform the data base to be list of features
- Convert the value to numerical
- Fill in all the NA with 0 (Considering filling with mean value instead)
- Data Normalization



Sample (unnormalized):

Residential Energy Consumption Prediction
vs price features (named in MSB index):

Predict next year(or specified year) energy
consumption based on other given prices.

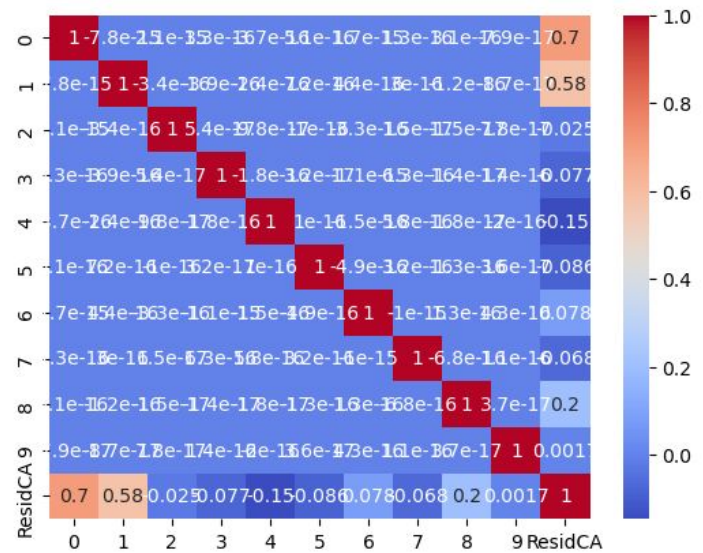
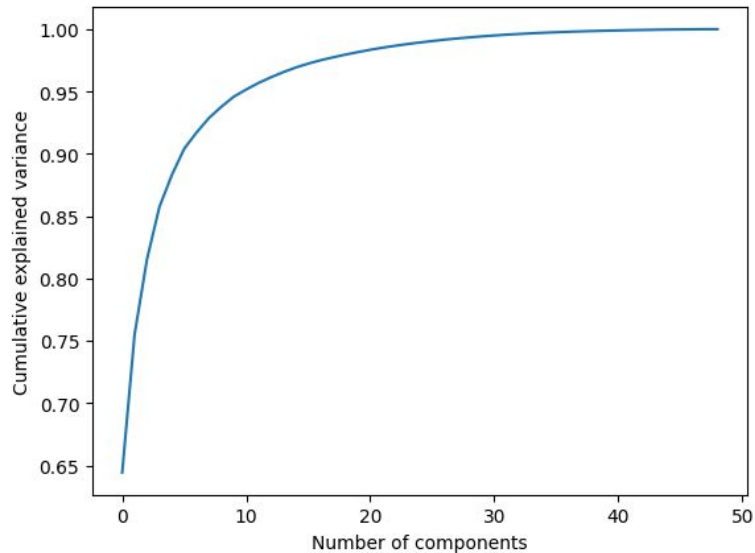
Years	ARICD	...	ZWCDP	Residenti al CA
1970	0.49	...	3169	1192848
...
2020	12.56	...	2539	1507721



Data Visualization

PCA analysis

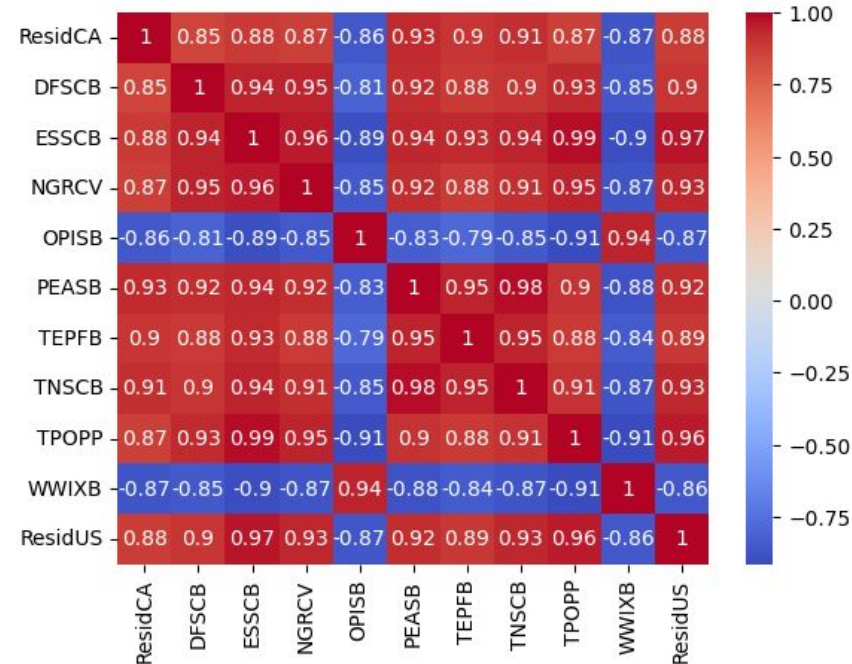
Select 10 as hyperparameter for pca analysis



Correlation analysis

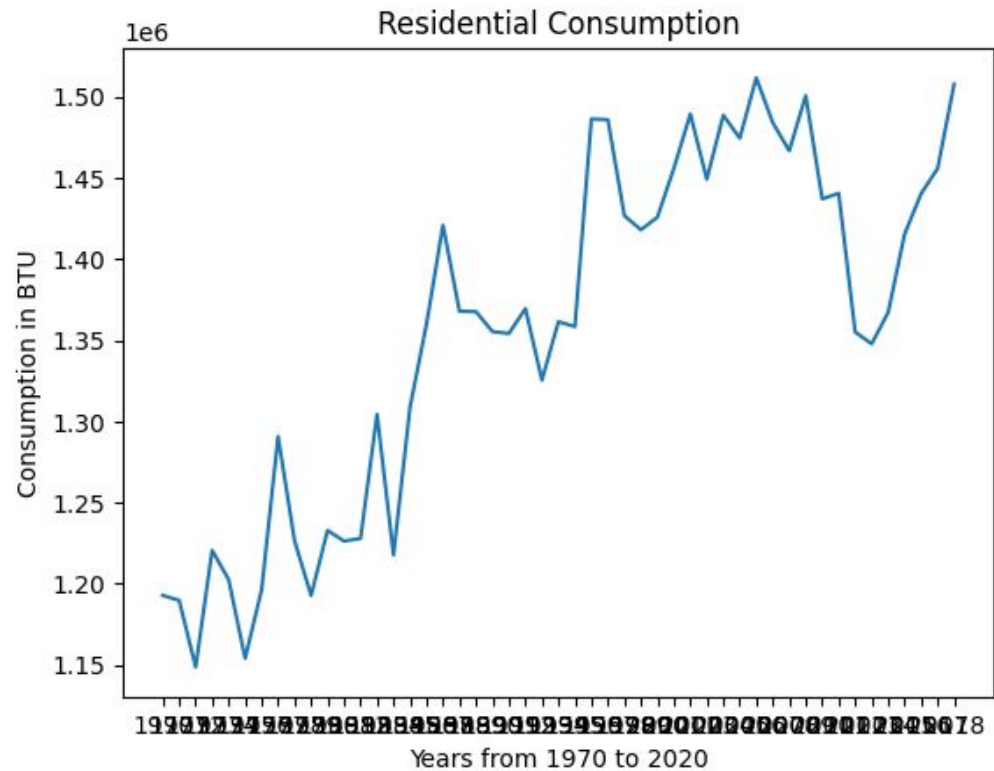
- Select features with high correlation with target (>0.85)
- Eliminate duplicate features:
TNASB & PEASB
OPSCB & OPISB
- Features from 306 down to 11

Two year Residential Sector Prediction Sample



Sample target

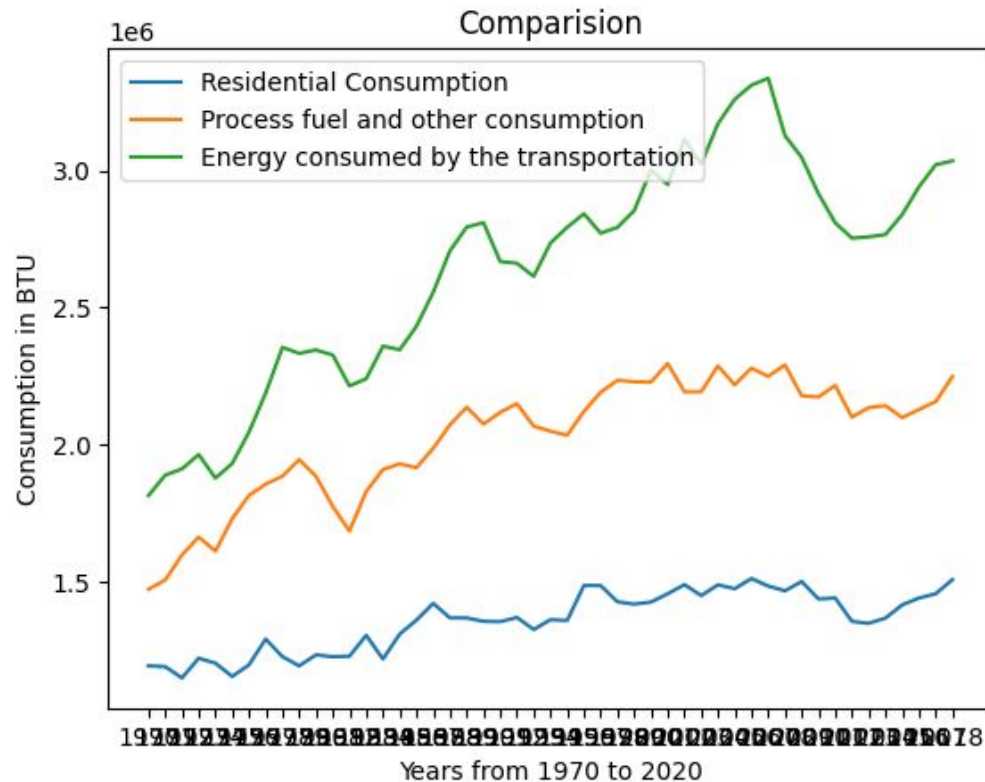
Residential Consumption in CA



Residential consumption growth over years.

Sample Relation

Residential Consumption in CA

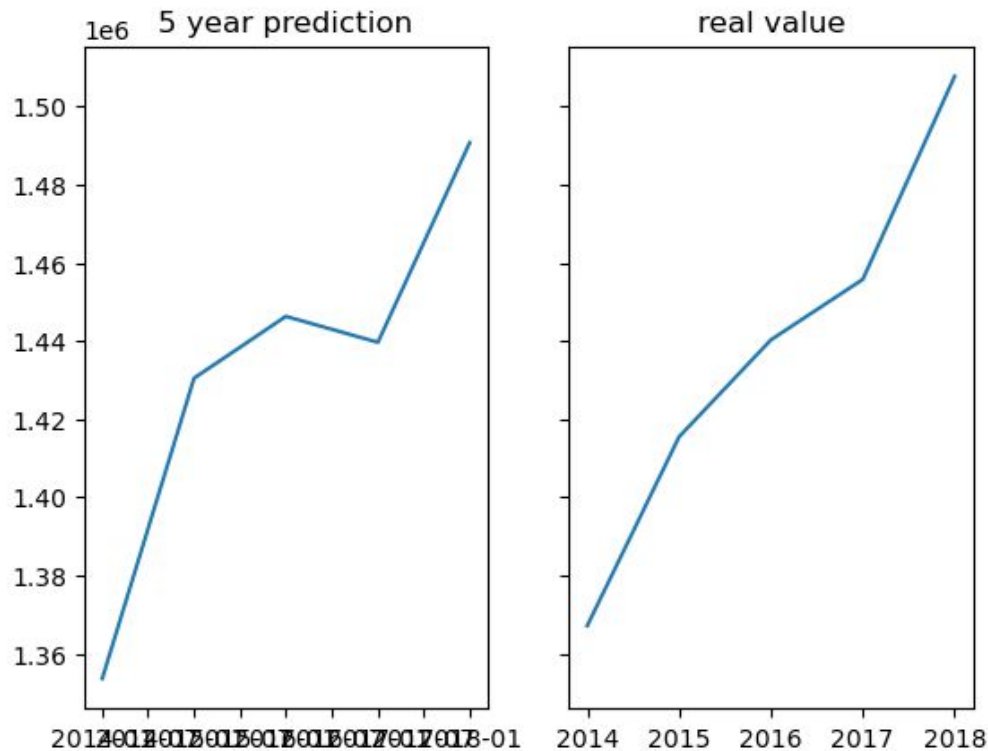


Residential consumption Compare with main features

SARIMAX Validation

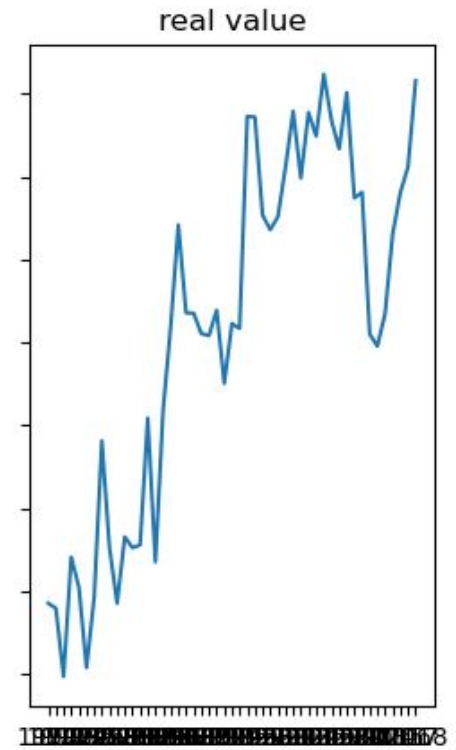
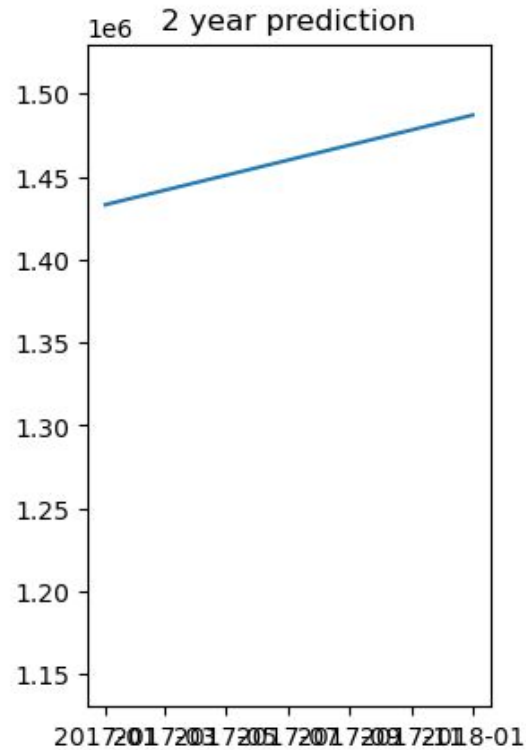
5 year prediction based on pca data

MAPE: 0.009399553213092346



SARIMAX Prediction

2 year prediction based on pca data





Milestone

We cleaned the data, wrangled data and partial analysis and modeling.

Now we are heading to dig deeper in analysis and apply other model for comparisons.

- Winter
 - Weeks 8-10 Data sets Collection and Combination (PostgreSQL database, github repo)
- Spring
 - Weeks 1-2 Data sets EDA (Method defined and validated, applying to other models)
 - Weeks 3-4 Statistic Model fit, and Machine learning Model fit (SARIMAX model applied, working on other models)
 - Weeks 5-6 Model validation and visualization (SARIMAX model applied, working on other models)
 - Weeks 7-8 Model analyze (SARIMAX model applied, working on other models)
 - Weeks 9-10 Final report