DSO 522 Project Report

Forecasting U.S. Power Company

Team Members
Yifeng Wang
Yun Guo

Instructor
Dr. Robertas Gabrys

Table of Contents

1. Executive Summary

The purpose of this project is to find the best model to predict hourly load of 2005 and daily peak hour of 2006 in the United States. According to our discussion and preliminary analysis, we applied the following forecasts:

- Naïve
- Time Series Decomposition
- ARIMA
- Holtz-Winter's Exponential Smoothing
- Multiple Linear Regression

Each model's advantages and its application purpose in our project.

2. Introductory Remarks

2.1. Introduction and Motivation

The dataset we have includes three years hourly load data from 2002 to 2004 and four years of hourly temperature data from 2002 to 2005, provided by U.S. Power company. According to the dataset, we expect to forecast the hourly load in 2005 and daily peak hours in 2006. The hourly load in 2005 will be evaluated by MAPE and RMSE. The daily peak hours will be evaluated by the total number of correct hours during 2006. Thus, we need to build more specific forecasting models to predict future hourly load and daily peak hours even though it is hard to capture all specific dynamics as our operational attributes or effects.

The following sections will give more detailed explanation about our models and analysis.

2.2. Hypothesis

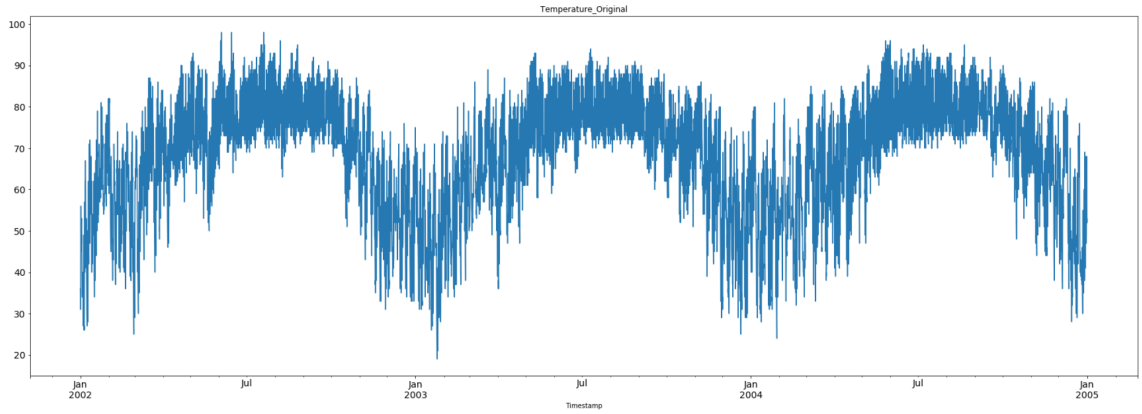
First, we assumed that the temperature (T) had effect on load. On the basis of common sense, we assumed that the absolute value of actual temperature and body temperature also had impact on load.

Second, we assumed that the seasonality of time influence on load. Because people use different amount of electricity in different seasons, particular in Summer or Winter.

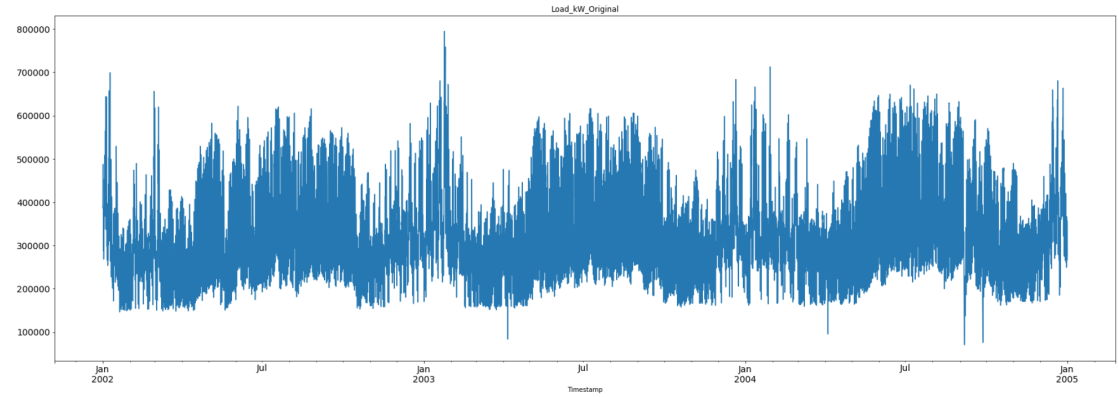
2.3. Data Exploration and Visualization

2.3.1. Original Data Visualization (Hourly)

Original Temperature Description:

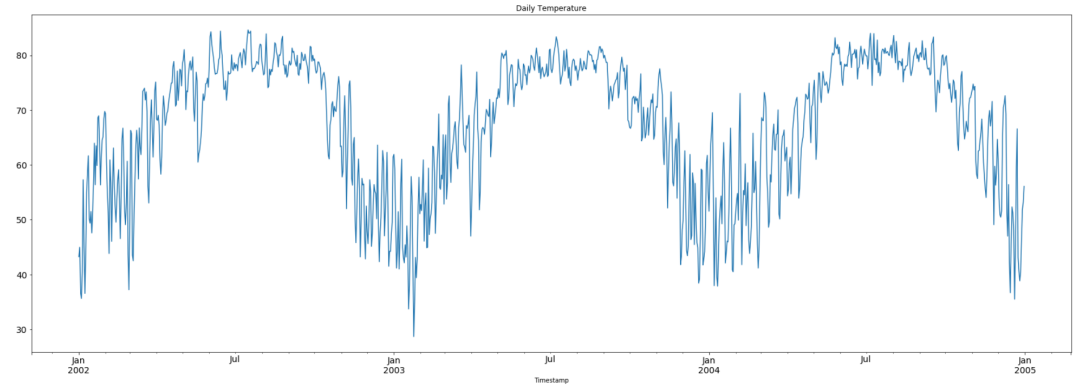


Original Load_kW Description:

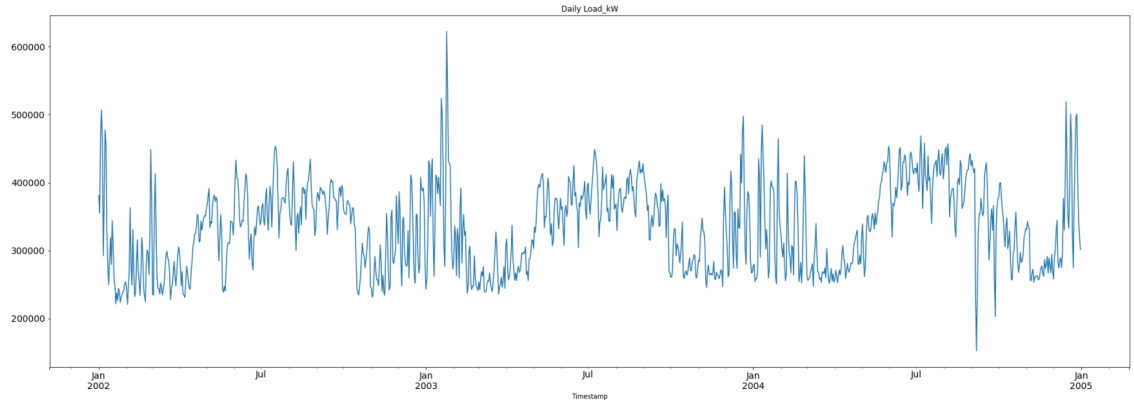


2.3.2. Daily Data Visualization

Daily Temperature Description:

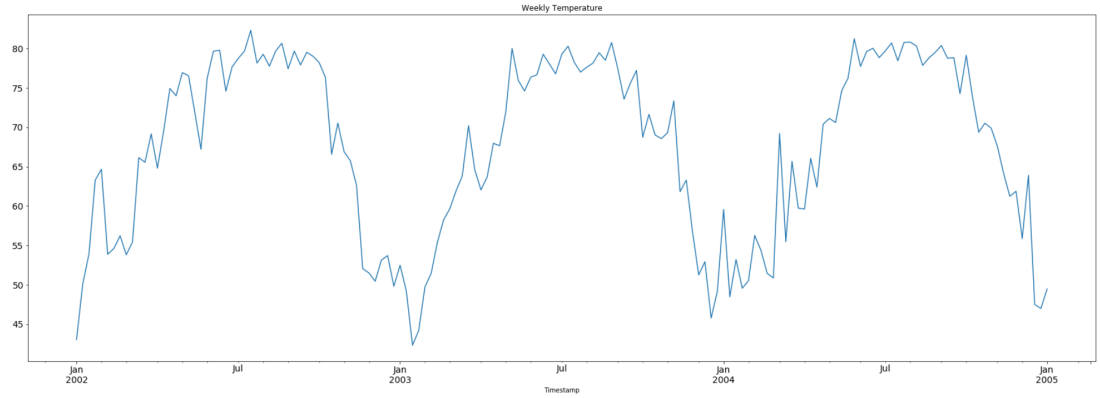


Daily Load_kW Description:

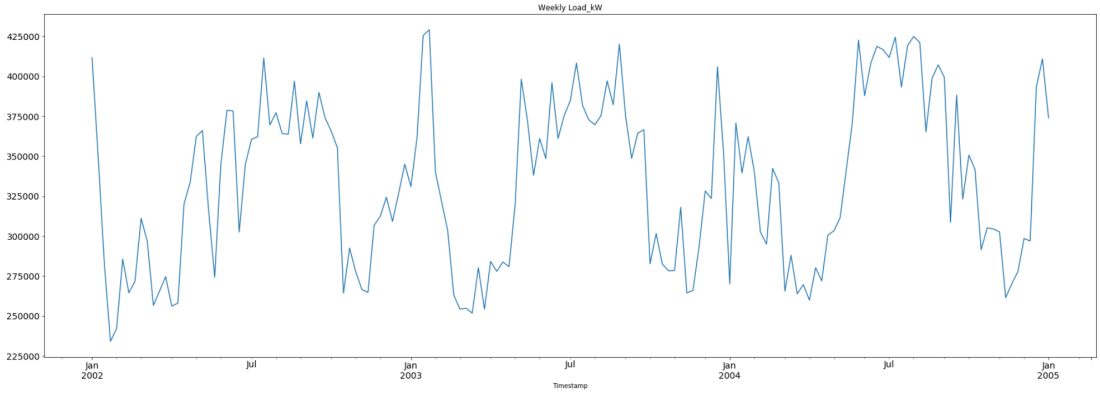


2.3.3. Weekly Data Visualization

Weekly Temperature Description:

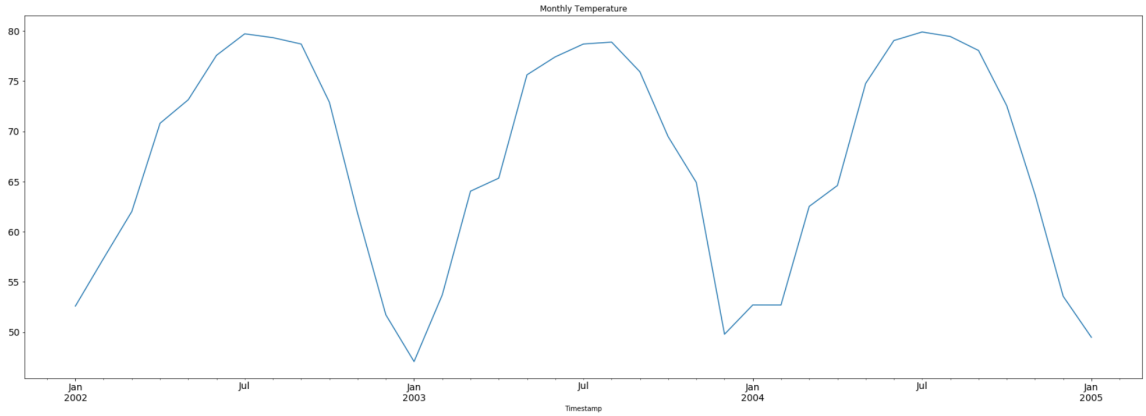


Weekly Load_kW Description:

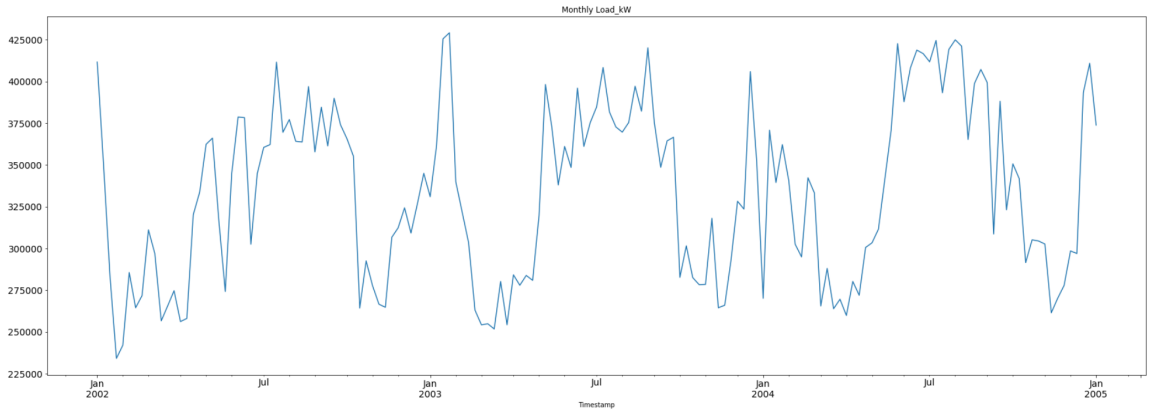


2.3.4. Monthly Data Visualization

Monthly Temperature Description:



Monthly Load_kW Description:



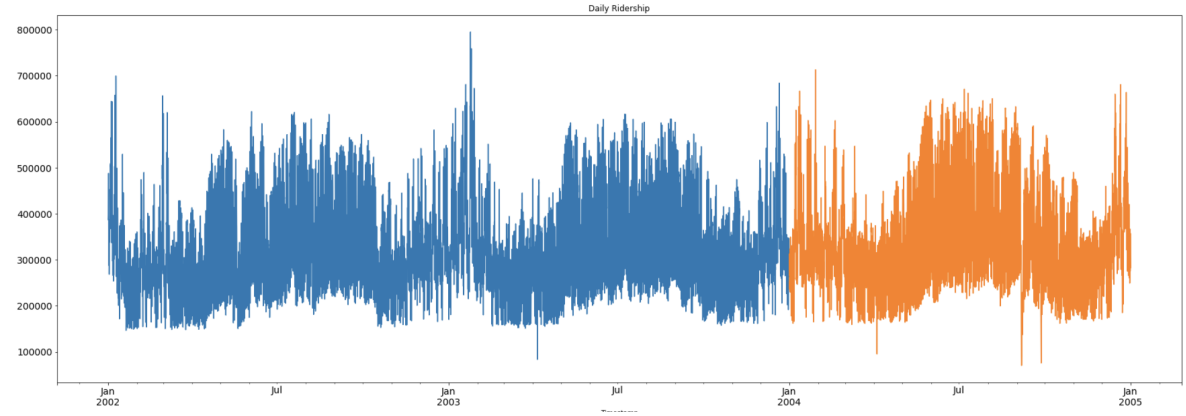
3. Forecasting Method

3.1. Training/ testing period for selecting the best model

Since we don't have data for Hourly loads of 2005. We will use data from January 2002 to December 2005 as our training/ testing data period to determine our final model for this competition.

The training data period we use is between Jan/2002 to Dec/2004.

The testing data period we use is between Jan/2005 to Dec/2005.



3.2. Naive

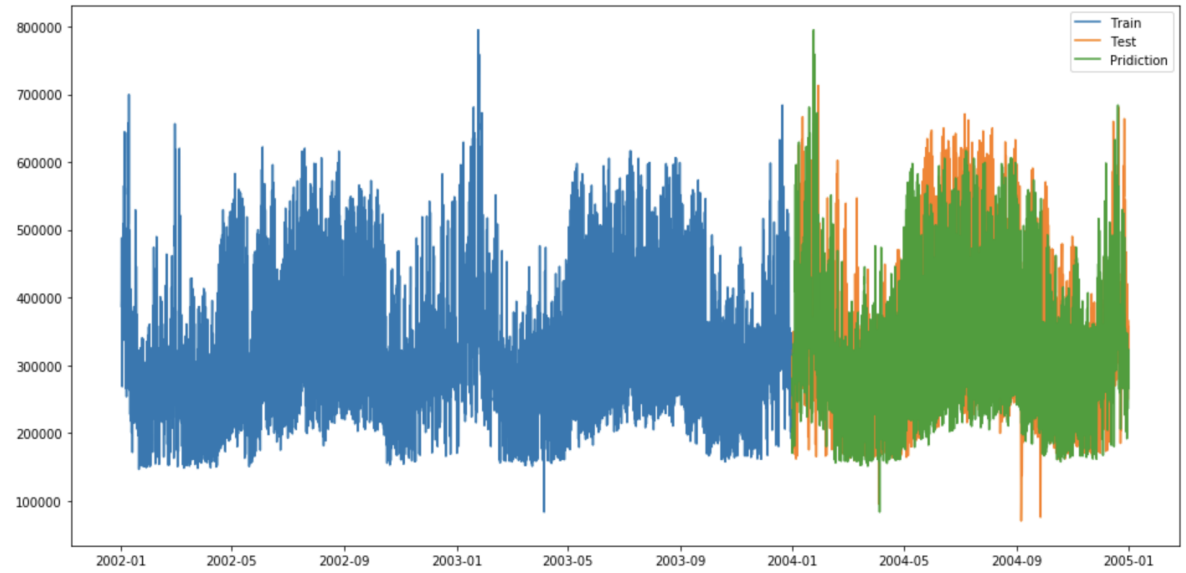
Definition: Estimating method in which the last period's actuals are used as this period's forecast, without adjusting them or attempting to establish causal factors. It is used only for comparison with the forecasts generated by the better (sophisticated) methods.

Model:

the training data is in blue color.

the testing data is in orange color.

the prediction data is in green color.

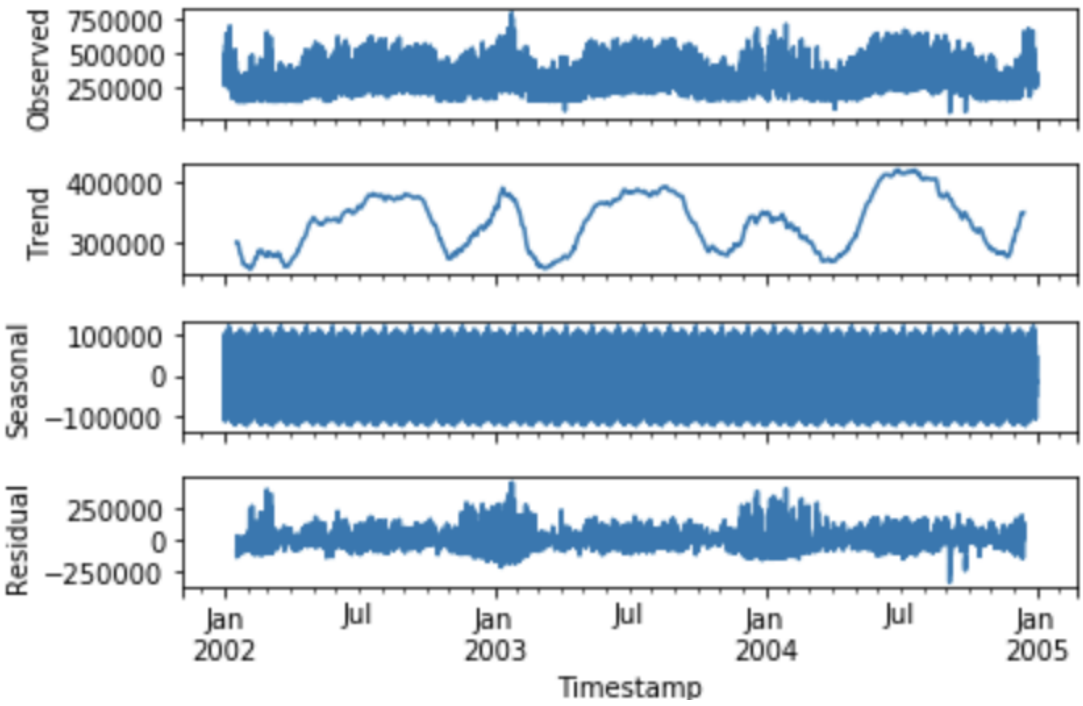


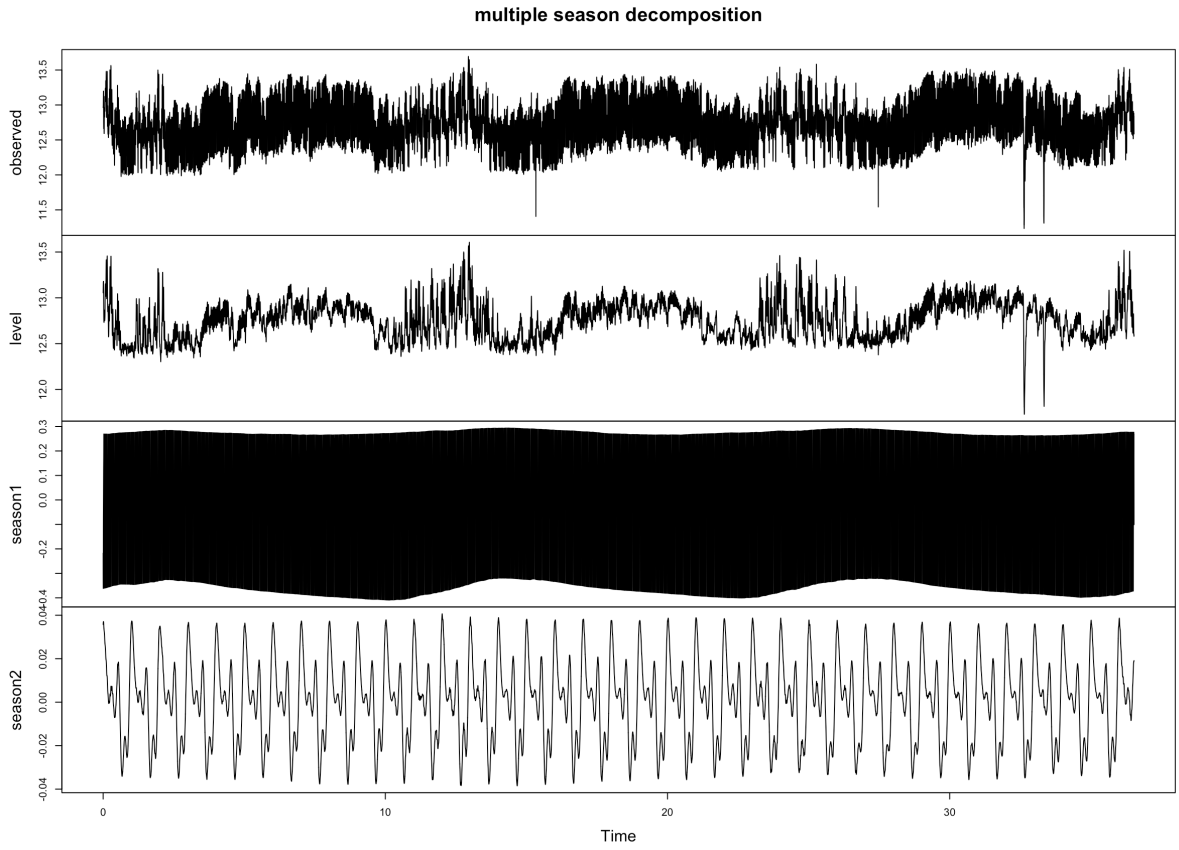
Outcome:
the RMSE for the test dataset is 83807.95.
The MAPE for the test dataset is 17.87%.

3.3. Time Series Decomposition

Definition: Decomposition methods are based on an analysis of the individual components of a time series. The strength of each component is estimated separately and then substituted into a model that explains the behavior of the time series. Two of the more important decomposition methods are Multiplicative Decomposition and Additive Decomposition. We use Additive Decomposition to forecast.

Model:

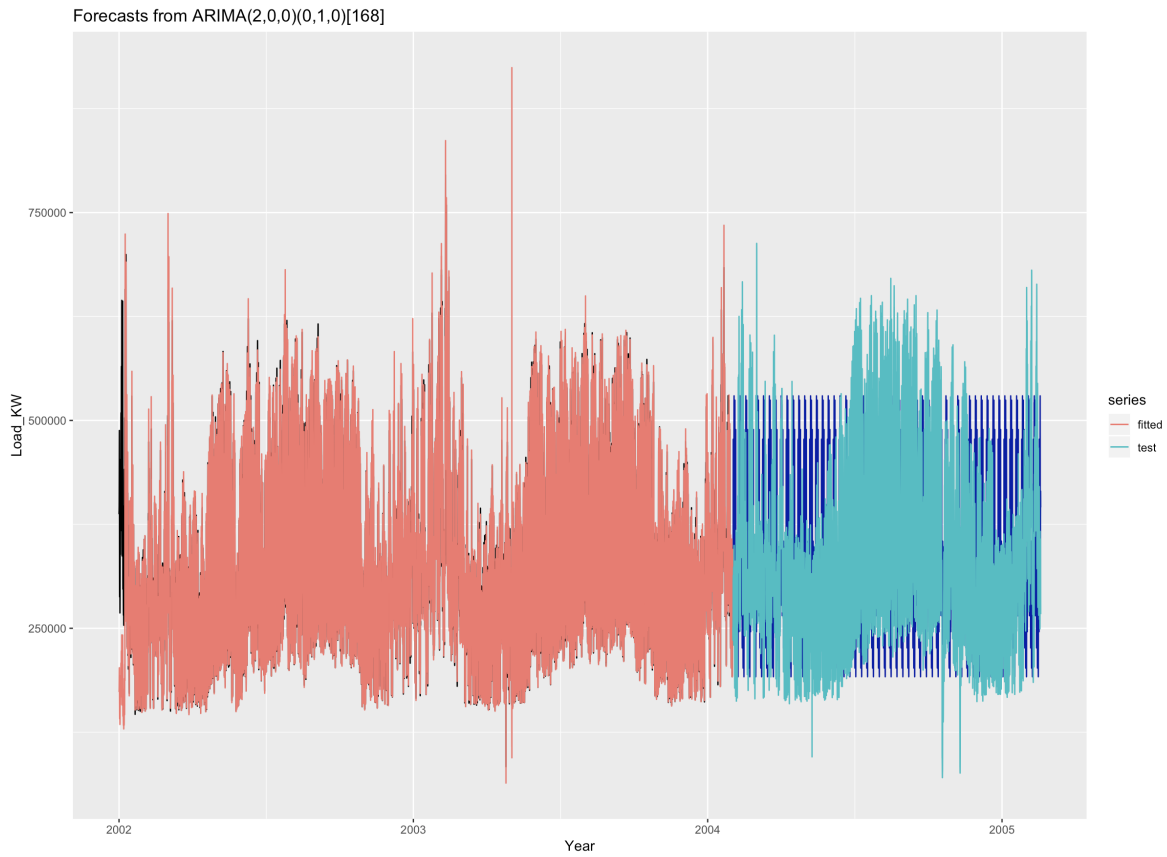




Outcome:
the RMSE for the test dataset is 95276.95.
The MAPE for the test dataset is 19.58%.

3.4. ARIMA

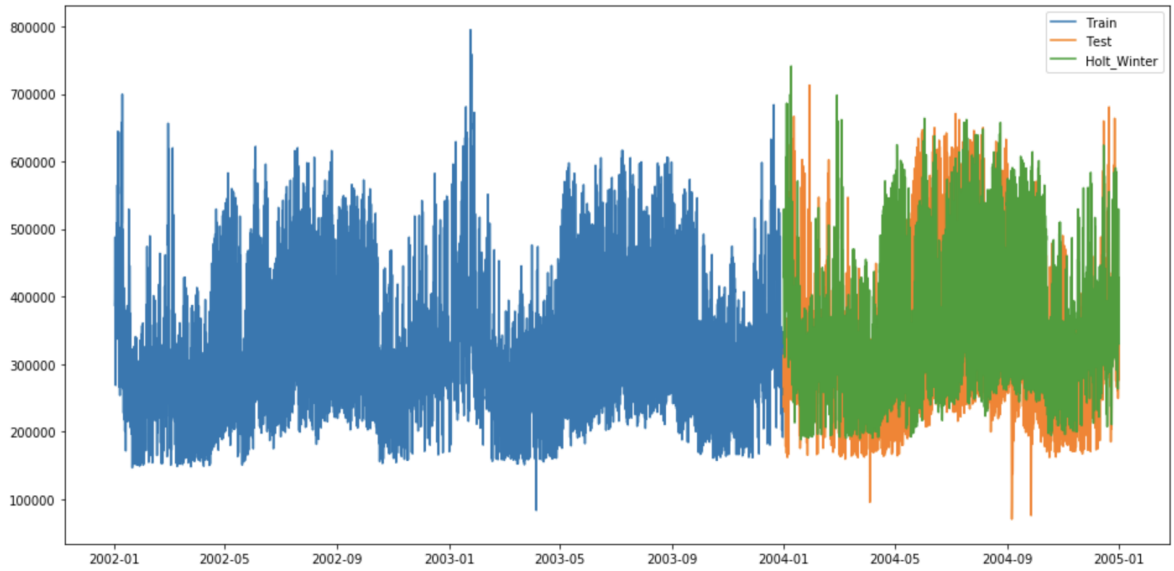
Definition: ARIMA model explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts. ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a generalization of the simpler AutoRegressive Moving Average and adds the notion of integration. This acronym is descriptive, capturing the key aspects of the model itself.
Model:



Outcome:
the RMSE for the test dataset is 136403.20.
The MAPE for the test dataset is 31.76%.

3.5. Holtz-Winter's Exponential Smoothing

Definition: Holt-Winters is a model of time series behavior. Forecasting always requires a model, and Holt-Winters is a way to model three aspects of the time series: a typical value (average), a slope (trend) over time, and a cyclical repeating pattern (seasonality). Holt-Winters uses exponential smoothing to encode lots of values from the past and use them to predict "typical" values for the present and future.
Model:



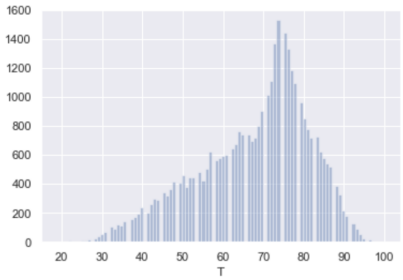
Outcome:
the RMSE for the test dataset is 91779.29.
The MAPE for the test dataset is 21.84%.

3.6. Linear Regression

Definition: Linear regression is a statistical tool used to help predict future values from past values. For time-series analysis, it is possible to develop a linear regression model that simply fits a line to the variable's historical performance and extrapolates that into the future. If the variable is plausibly linear, using linear regression to forecast might yield a useful prediction.

Operation: There are several following variables (attributes) which influence on the hourly load forecast we considered:

1) Actual Temperature: 1 variable
2) The absolute value of the Actual Temperature and the Body Temperature (We assumed that the most comfortable body temperature is most common temperature, which is 74 degrees Fahrenheit.): 1 variable



3) Current Time (We divided the 24 hours in a day into 12 time periods, two hours into one time period, and one week has 7 days): 12 * 7 = 84 variables.
4) Combine Current Time with Temperature as variables (Multiplicative relationship): 12 * 7 * 1 = 84 variables
5) Combine Current Time with the absolute value of the Actual Temperature and the Body temperature as variables (Multiplicative relationship): 12 * 7 * 1 = 84 variables

The total of variables we considered are 1 + 1 + 84 + 84 + 84 = 254 variables. Based on these variables, we could get the best MAPE and RMSE value, which is the best model for our forecasting.
Model:

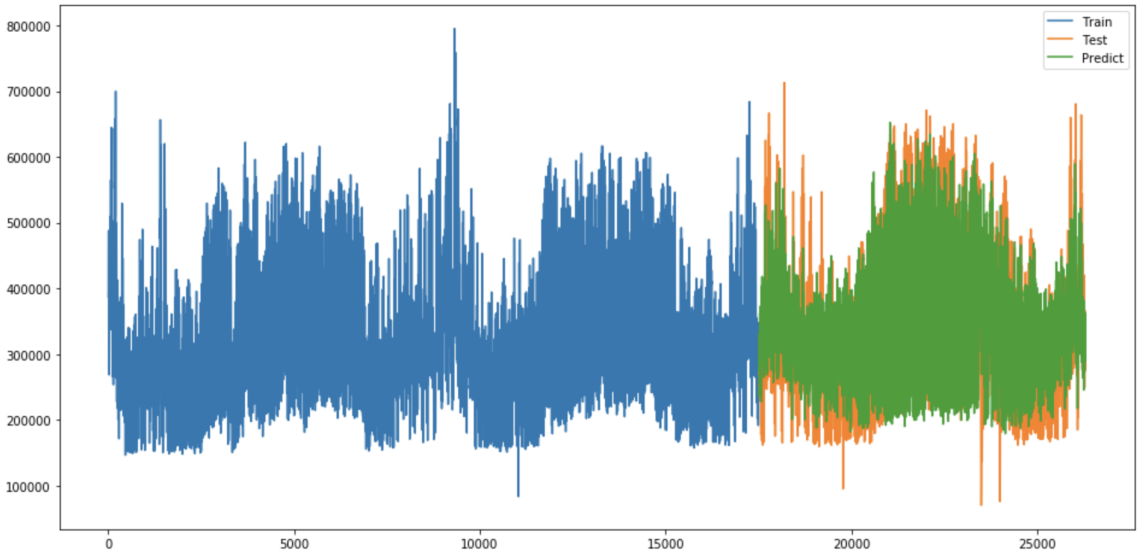
Outcome:
the RMSE for the test dataset is 63400.78
The MAPE for the test dataset is 14.27%.

3.7. Ensemble model(linear regression and naive)

Definition:
Ensemble modeling is the process of running two or more related but different analytical **models** and then synthesizing the results into a single score or spread in order to improve the accuracy of predictive analytics and data mining applications. In this case, we choose our two best models, linear regression and naive model.

y_ensemble = 0.6 * y_linear_regression + 0.4 * y_naive

Model



Outcome:
the RMSE for the test dataset is 59256.41.
The MAPE for the test dataset is 12.88%.

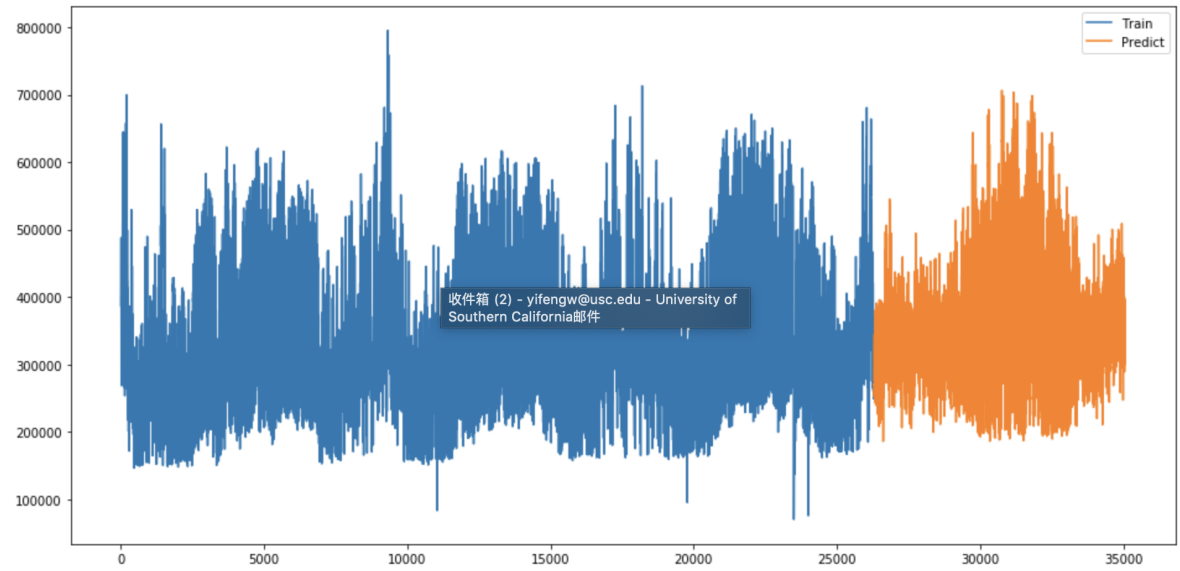
3.8. Summary Table

Model	RMSE	MAPE
Naive	83807.95	17.87%
Time Series Decomposition	95276.95	19.58%
ARIMA	136403.20	31.86%
Holtz-Winter's Exponential Smoothing	91779.29	21.84%
Linear Regression	63400.78	14.27%
Ensemble model (linear regression + naive)	59256.41	12.88%

4. Final (best) model for this competition

4.1. Forecasting Hourly loads of 2005

We use ensemble model(linear regression and naive) to do the prediction.
We use data from Jan 2002 to December 2004 to train the model.
Then we use this ensemble to predict the Hourly loads of 2005.



4.2. Forecasting Daily peak hours of 2006

We use Holtz-Winter's Exponential Smoothing to do 2-year prediction.
Then we extract the daily peak hours from our prediction data.

5. Conclusion

Given the data provided, we first build several models to calculate the RMSE and MAPE, using the training set of 2002, 2003 and testing set of 2004. Then we choose the best RMSE of those models, use the linear regression model and ensemble model to forecast the hourly load. The linear regression model will be used to get the prediction of 2005, since it considers the temperature of 2005, as well as the time component. The ensemble model will be used to forecast the hourly load of 2006, using the data of 2002-2004, to predict the load of 2005 and 2006, then we will select the daily pick hours in those hourly load.