Forecasting Average Temperature in California

Phillip Jauregui
Department of Applied Analytics, Wood College of Advancing Studies, Boston College

Abstract

This poster focuses on forecasting average temperature in California using four models: seasonal naïve (SNAIVE), ARIMA, ETS, and ensemble. The analysis emphasizes the importance of studying temperature trends in regions vulnerable to climate change. SNAIVE serves as a baseline, while ARIMA captures autoregression, differencing, and moving average components, and ETS models error, trend, and seasonality. Ensemble models combine multiple models to enhance prediction accuracy.

The data reveals clear seasonality and no significant upward trend. The ensemble model outperforms others on the test set, with lower MAPE and MAE values, potentially due to its ability to handle complex seasonality, non-linear trends, and outliers. While the ensemble model excels in this dataset, other models remain valuable for different forecasting time frames. Choosing the best model depends on data characteristics and the problem at hand. Ensembles' strength lies in combining models to produce accurate and robust predictions, making them advantageous in specific scenarios such as predicting average temperature on a yearly basis.

Introduction & Significance

The Problem:

As climate change continues certain areas will be hit worse than others (Hsiang et al., 2017). Accordingly, examining the average temperature over time in places likely to be affected by climate change becomes increasingly important.

Methods Employed:

Temperature data is here analyzed using seasonal naïve, ARIMA, ETS, and ensemble models. Seasonal naïve forecasting is simple yet useful when data shows repeating patterns (i.e., seasonality). It predicts future values based on the past observations from the same season in previous years. While easy to use, it may not consider other factors and is best suited as a baseline for more advanced forecasting models (Hyndman & Athanasopoulos, 2018). One such more advanced model type is ARIMA, which combines autoregression (AR) to capture past values' influence, differencing (I) to make data stationary, and moving averages (MA) to consider past forecast errors (Box et al., 2015). ETS modeling is a time series forecasting method used to capture error, trend, and seasonality components in data. Lastly, ensemble models combine multiple individual models to improve prediction accuracy (Wu & Levinson, 2021).

Significance of this Analysis:

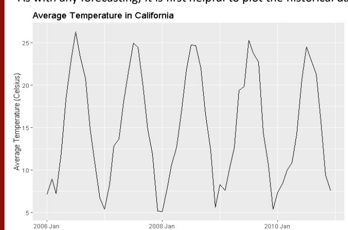
Analyzing temperature data over time allows us to gain insights into climate change patterns and trends. In regions like California, which are prone to wildfires, temperature forecasts play a crucial role in assessing the risk of fire outbreaks and planning firefighting efforts and evacuation measures. Furthermore, Weather forecasts are vital for decision-making across industries like agriculture, energy, tourism, and transportation. Accurate temperature forecasts enable farmers to plan crop activities, aid energy companies in predicting demand, and help tourists plan vacations.

Goal of this Analysis:

Offer precise forecasts to aid in informed decision-making, risk mitigation, enhancing overall resilience to and awareness of weather and climate changes.

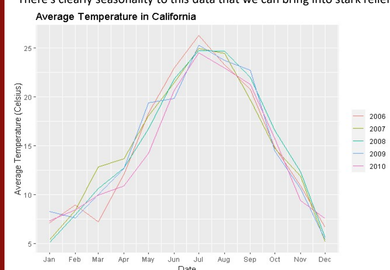
Methods

As with any forecasting, it is first helpful to plot the historical data:



Plot 1. Average Temperature in California from 2006 – 2010.

There's clearly seasonality to this data that we can bring into stark relief:



Plot 2. Average Temperature in California from 2006 – 2010, broken up by month

Reasonably, temperatures rise seasonally in the summer and drop in the winter. Given the short time frame of the data, we do not see any upward trend in overall average temperature, but this does not necessarily signify a lack of Increasing overall temperatures. Instead, more historical data should be considered to explore this possibility.

To forecast the temperature data, I first used 80% of the data as a training set to build 4 models: seasonal naïve (SNAIVE), ARIMA, ETS, and ensemble.

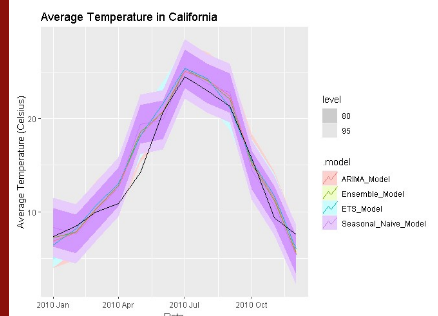
- **SNAIVE**:
  - relies on the most recent observation from the same season in the previous year as the prediction. Most suitable for data with strong seasonality and as a baseline comparison with more advanced models.
- **ARIMA** (AutoRegressive Integrated Moving Average):
  - Captures autoregression, differencing, and moving average components to model time series data. Effective for short to medium-term forecasting and handling data with trends and seasonality.
- **ETS** (Error-Trend-Seasonality):
  - Decomposes data into error, trend, and seasonality components. Like ARIMA, ETS is useful when dealing with data exhibiting both trend and seasonality.
- **Ensemble**:
  - Combines multiple individual models to produce more accurate and robust predictions. Effective for improving accuracy and generalization.

Results

Estimating the models on the withheld 20% test set, we get a sense of which forecasting model performed best.
Specifically, comparing across models, we see that the Ensemble model has the least error:

	SNAIVE	ARIMA	ETS	Ensemble
RMSE	1.91	1.40	**1.03***	1.61
MAPE	12.04	12.04	10.91	**10.67***
MAE	1.45	1.45	1.27	**1.23***

Table 1. Accuracy metricsfor the 4 forecasting models

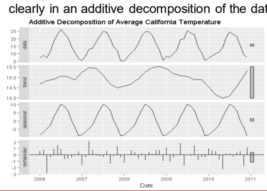


Plot 3. Joint forecast models for the 20% test set of 2010 temperature data.

Looking at the plot alone, it is difficult to discern which model best fits our data. Thankfully, we can look at accuracy metrics like to determine which forecasting model performs best.

RMSE (Root Mean Squared Error) evaluates the accuracy of a predictive model by measuring the average difference between predicted values and the actual observed values in a dataset. Accordingly, we prefer models that minimize RMSE, thus exhibiting less error in their predictions.
Similarly, we say that models with lower MAPEs and lower MAEs perform better than those with high values on these metrics.

Applied to the current case, we see that the ETS model has the lowest RMSE. But, the ensemble model outperforms all other models on MAPE and MAE. Accordingly, the ensemble model best fits our data and is our preferred model with the current dataset. One potential reason that the data displays non-linear trends, which ARIMA assumes. We can see this more clearly in an additive decomposition of the data:



Plot 4. Additive decomposition of the full dataset (2006-2010)

Discussion

Model Performance:
All models performed well in forecasting the withheld 20% test set.
The Ensemble model, however, outperformed all others, as seen by the lowest MAPE and MAE values. However, it is noteworthy that the ETS model had the lowest RMSE and is quite close to the ensemble model in fitting the data.
The ensemble model could have outperformed the other models for a number of reasons.

1. Complex Seasonality: the data displays strong seasonality that may not be well-captured by SNAIVE or ARIMA models.
2. Non-Linear Trends: ARIMA models assume linear trends, and we can see from the additive decomposition plot that the data here displays a non-linear trend.
3. Robustness to Outliers: Ensemble (and ETS) models can handle outliers and extreme values by incorporating them into the error component, while ARIMA models might be sensitive to outliers and produce less accurate predictions.
4. Reduced Bias: Individual models, like SNAIVE, ETS, or ARIMA, may suffer from inherent biases or assumptions that limit their ability to accurately capture specific patterns in the data. Ensembles, by aggregating predictions from multiple models, can reduce the overall bias and provide more balanced and accurate forecasts.

Notably, the best model depends on the specific characteristics of the data, the problem at hand, and the modeling assumptions. As such, it is a best practice to simultaneously fit multiple models and select that which performs best based on accuracy metrics like RMSE, MAPE, MAE, etc. Ensemble models, in particular, can be advantageous in situations where combining different models' strengths leads to superior predictions, even if individual models like ETS show good performance.

Conclusion

An Ensemble model built on 4 years of training set data best forecast 1 year of test set data regarding average temperature in California. The Ensemble model may have outperformed the other models (SNAIVE, ARIMA, ETS) due to better capturing the seasonality of the data as well as the data's non-linear trend while also better handling outliers.
Nevertheless, the other models are important to forecast especially for forecasts of varying time frames. In the current situation where the plot does not display a clear best performing model, accuracy metrics become paramount in evaluating model fit.

References

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015*). Time series analysis: Forecasting and control* (5th ed). John Wiley & Sons.
Hyndman, Rob J., and George Athanasopoulos. *Forecasting: principles and practice* (3rd Edition). OTexts, 2018.
Solomon Hsiang et al. ,Estimating economic damage from climate change in the United States. *Science* 356, 1362-1369 (2017). DOI:10.1126/science.aal4369
Wu, H., & Levinson, D. (2021). The ensemble approach to forecasting: a review and synthesis. *Transportation Research Part C: Emerging Technologies*, *132*, 103357.

BOSTON COLLEGE
Woods College of Advancing Studies