Extended Abstract

Title: Foodborne Illness Time Series Forecasting and Analysis using ARIMA, Prophet, and Machine Learning

Introduction:

Foodborne illness outbreaks are a continual public health hazard from time to time in the U.S., thus needing timely forecast for authority preparedness and intervention. This project uses statistical modeling and machine learning models to analyze and forecast monthly outbreak trends through estimating effects of temperatures with CDC outbreak data from 1998-2015.

Methods:

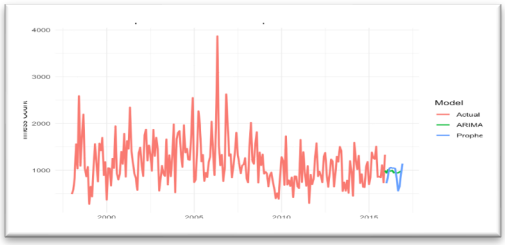
We used various packages such as packages, prophet, changepoint, tidymodels and randomforest. We began with data preprocessing where we remove the zero-illness records to focus solely on outbreak-related events. The monthly aggregate count of illnesses was analyzed for trend and seasonality using ARIMA and Prophet time series models. We identified structural breaks with the PELT algorithm for changepoint detection. To enhance prediction granularity, segment-wise ARIMA models were fit to each stable interval. In parallel, a Random Forest regression model was trained on engineered features including year, month, state, location type, species, and outbreak status. Model evaluation incorporated multiple metrics and classification performance based on outbreak thresholds.

Results Highlights:

The ARIMA model trained on the full time series achieved a Root Mean Square Error (RMSE) of 269.4 and a Mean Absolute Error (MAE) of 237.0, while the Prophet model yielded an RMSE of 281.2 and an MAE of 231.0. A comparison of predicted peak months revealed that ARIMA provided earlier warnings, predicting the peak 3 months ahead, whereas Prophet predicted it 1 month early—both supporting timely public health response. Ensemble forecasting, created by averaging ARIMA and Prophet outputs, further improved accuracy with an RMSE of 260.2 and MAE of 203.2. For outbreak classification, both models achieved 83.3% accuracy and perfect sensitivity, although specificity remained low.

Model Output:

Model	RMSE	MAE
<chr>	<dbl>	<dbl>
ARIMA	269.4506	236.6162
Ensemble	260.0899	203.4052
Prophet	280.7426	230.5424



Conclusion:

Both time series and machine learning models generated meaningful insights into forecasting foodborne illness. The ARIMA and Prophet models identified trends and seasonality, whereas the Random Forest uncovered contextual drivers. The ensemble methods showed promise in increasing predictive accuracy. The early detection timeline (lead time) indicated the utility of these methods for public health planning