Predicting COVID-19 Deaths (Jan 2020 - Jan 2024)

Final Presentation

I <3 DS

Alissa Chu, Ryan Nguyen, Nishi Shah, Emily Zhang

Presentation Overview

1. Problem Introduction
2. Data Processing & Feature Selection/Engineering
3. ARIMA Models
    a. ARIMA
    b. SARIMA
4. Prophet Models
5. XGBoost
6. Analysis + Conclusions

Problem Introduction

- **Objective**: Analyze COVID-19 vaccination data in order to predict COVID-19 death rates within 4 US regions: East, Midwest, South, and West, with different temporal focuses
- **Univariate Models**: Death data was daily. We sought to predict daily COVID-19 death rates with three univariate models: ARIMA, auto-ARIMA/SARIMA, and Prophet (1136 rows)
- **Multivariate Models:** Death data was weekly. We sought to predict weekly COVID-19 death rates with two multivariate models: Prophet and XGBoost. (7306 rows, 91 columns)
- **Goal:** Provide a robust analysis of the relationship between vaccination rates and seasonality with covid deaths and assist vaccine manufacturers, health professionals, and government officials to **prevent future COVID-19 deaths**

Data Processing & Feature Selection/Engineering

- **EDA & Data Preparation**
  - Cleaned & merged vaccine & death data set for multivariate models
  - Removed redundant columns
  - Aggregated state data into regions: East, Midwest, South, West
- **Feature Engineering:**
  - Added lag and rolling window statistic using half year, year, year and a half, and two years as our lags/windows
  - Created 'region' column based on US state, a 'season' column based on month, and a 'holiday' column -> one hot encoded those
- **Feature Selection:**
  - Created a **selected features dataset** to be used for our multivariate models:
    - Preliminary XGboost model with all of the columns in our weekly deaths dataset (our merged vaccine and weekly deaths)
    - Selected features from the optimal XGBoost model with a feature importance above 0.005 and were left with 65 selected features
    - Added our 26 lag features and rolling window statistics
  - Total of 91 columns in our multivariate selected features dataset

Additional Data Prep

Univariate (ARIMA models) Differencing

- Applied differencing for non-stationary regions (Midwest, South, and West)
- Before differencing

East	Midwest	South	West
ADF Statistic: -3.842138	ADF Statistic: -2.473716	ADF Statistic: -3.104854	ADF Statistic: -2.816341
p-value: 0.002504	p-value: 0.122001	p-value: 0.066196	p-value: 0.055977

- After differencing

East	Midwest	South	West
ADF Statistic: -3.842138	ADF Statistic: -6.612100	ADF Statistic: -6.157430	ADF Statistic: -5.691688
p-value: 0.002504	p-value: 0.000000	p-value: 0.000000	p-value: 0.000001

Multivariate (Prophet)

- Imputed 1,000,000 for missing vaccination data to empower the models to discern underlying patterns within the data

Metrics

MAE

- Optimal metric for cross-model comparison
- Used for comparison of our Multivariate models

MASE

- Optimal metric for time-series regression because it is a scaled version of MAE
  - Used for comparison of our Univariate models
  - Univariate looks at the different regions and MASE handles differences in datasets better
- Chose MASE over MAPE, because MAPE is sensitive to test values close to zero

Univariate Models

ARIMA Models

ARIMA

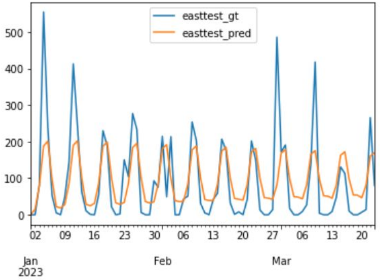
EAST	parameters	MAE	MASE
Alissa	6, 0, 6	59.725	0.674
Emily	5, 1, 6	57.1463	0.5981
Nishi	5, 1, 4	48.453	0.622
Ryan	5, 0, 5	47.188	0.724

MIDWEST	parameters	MAE	MASE
Alissa	6, 1, 6	60.029	0.541
Emily	6, 1, 6	61.101	0.532
Nishi	5, 1, 4	125.948	0.565
Ryan	5, 1, 5	54.436	0.515

SOUTH	parameters	MAE	MASE
Alissa	6, 1, 6	91.638	0.573
Emily	6, 1, 6	135.654	0.628
Nishi	5, 1, 4	131.212	0.530
Ryan	5, 1, 5	94.708	0.548

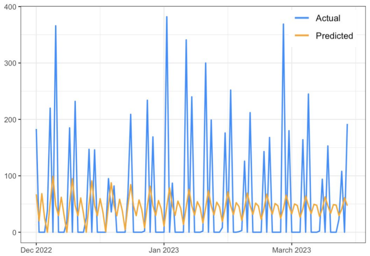
WEST	parameters	MAE	MASE
Alissa	6, 1, 6	36.165	0.451
Emily	6, 1, 6	37.028	0.449
Nishi	5, 1, 4	72.550	0.616
Ryan	5, 1, 5	46.880	0.594

BEST ARIMA MODELS



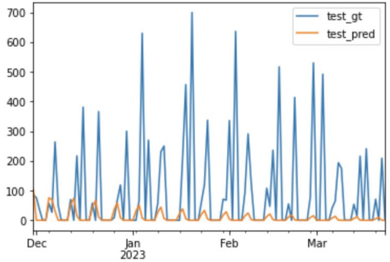
EAST REGION

- Used (d = 1)
- **Tuning**
- Grid search in range (1,7)
  - Params (5, 1, 6)



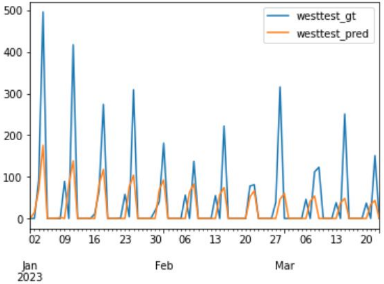
MIDWEST REGION

- Applied one difference to make data stationary according to ADF test (original p-value of 0.325)
- Grid search in range (0, 5)
  - Params (5, 1, 5)
- **Tuning**
- Clip test predictions AND test to be above zero since stationarity resulted in negative deaths



SOUTH REGION

- Applied one difference to make data stationary according to ADF test (original p-value of 0.302)
- **Tuning:**
- Grid search in range (1,3)
  - Params (5,1,4)
- Clip test predictions AND test to be above zero since stationarity resulted in negative deaths



WEST REGION

- Applied one difference to make data stationary according to ADF test (original p-value of 0.409)
- **Tuning:**
- Grid search in range (1,7)
  - Params (6,1,6)
- Clip test predictions AND test to be above zero since stationarity resulted in negative deaths

SARIMA

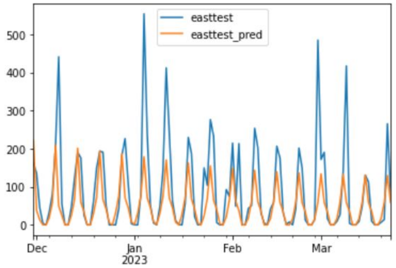
EAST	parameters	MAE	MASE
Alissa	(2,0 2)(2,0,2)[7]	52.947	0.597
Emily	(2,0,2)(2,0,2)[7]	50.0504	0.565
Nishi	(5,0,1)(2,0,2)[7]	73.042	0.622
Ryan	(5,0,4)(1,1,1)[7]	46.104	0.716

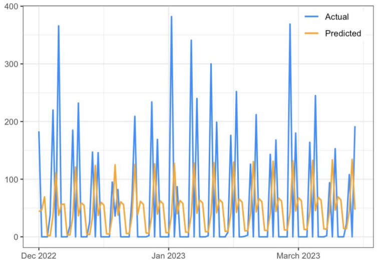
MIDWEST	parameters	MAE	MASE
Alissa	(1,1,2)(1,1,1)[7]	52.602	0.474
Emily	(1,1,2)(1,1,1)[7]	52.955	0.477
Nishi	(3,0,2)(1,0,1)[7]	84.702	0.565
Ryan	(0,1,3)(1,1,1)[7]	47.448	0.435

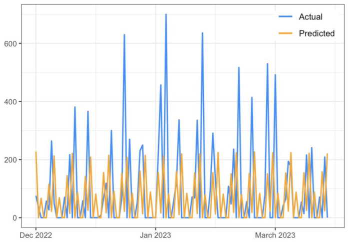
SOUTH	parameters	MAE	MASE
Alissa	(2,1,2)(1,1,2)[7]	85.775	0.536
Emily	(2,0,1)(2,0,2)[7]	98.629	0.510
Nishi	(5,0,2)(2,0,2)[7]	147.473	0.533
Ryan	(5,1,5)(1,1,1)[7]	85.738	0.496

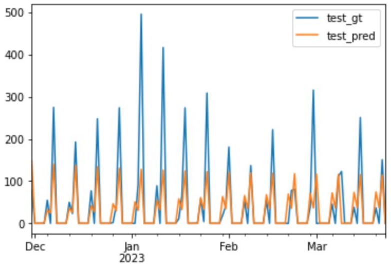
WEST	parameters	MAE	MASE
Alissa	(1,1,2)(2,1,2)[7]	29.869	0.373
Emily	(4,1,0)(2,1,0)[12]	33.100	0.417
Nishi	(4,0,1)(1,0,2)[7]	69.751	0.616
Ryan	(5,1,5)(1,0,1)[7]	36.964	0.456

BEST SARIMA MODELS









EAST REGION

- Already stationary (d = 0, D = 0)
- **Tuning:**
- Gridsearch in range (1, 3)
- Fit model:
  - order=(2, 0, 2)
  - seasonal_order=(2, 0,2,7)
- Clip test predictions AND test to be above zero
- Adjusted train: shifted vertically down 30 units

MIDWEST REGION

- Applied one difference to make data stationary according to ADF test (original p-value of 0.325)
- **Tuning:**
- Gridsearch in range (0, 5)
- Fit model:
  - order = (0, 1, 3)
  - seasonal_order=(1,1 ,1,7)
- Clip test predictions AND test to be above zero

SOUTH REGION

- Applied one difference to make data stationary according to ADF test (original p-value of 0.302)
- **Tuning:**
- Grid search in range (0,5)
- Fit Model
  - Order = (5, 1, 5)
  - seasonal_order = (1, 0, 1, 7)
- Clip test predictions AND test to be above zero since stationarity resulted in negative deaths

WEST REGION

- Applied one difference to make data stationary according to ADF test (original p-value of 0.409)
- **Tuning**:
- Grid search in range (1,3)
- Fit model:
  - order=(1, 1, 2)
  - seasonal_order=(2, 1,2,7)
- Clip test predictions AND test to be above zero

Prophet - Univariate

Univariate Prophet

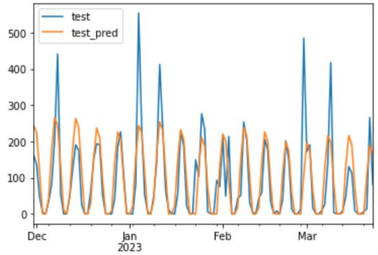
EAST	MAE	MASE
Alissa	42.219	0.476
Emily	57.2004	0.5987
Nishi	66.281	0.9994
Ryan	64.132	0.724

MIDWEST	MAE	MASE
Alissa	51.291	0.443
Emily	71.0213	0.607
Nishi	110.562	1.001
Ryan	92.441	0.798

SOUTH	MAE	MASE
Alissa	105.05	0.543
Emily	132.388	0.612
Nishi	157.526	0.999
Ryan	173.927	0.899

WEST	MAE	MASE
Alissa	42.053	0.470
Emily	68.016	0.733
Nishi	76.569	1.009
Ryan	74.746	0.835

BEST Univariate Prophet Models



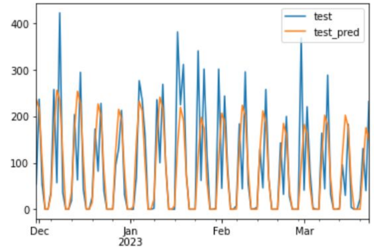
East

Tuned Model

- Changepoint_prior_scale = 0.02
- yearly_seasonality = False
- .add_seasonality(name='weekly', period=7, fourier_order=1)
- .add_seasonality(name='monthly', period=30.5, fourier_order=1)

Test metrics

- Clipped test predictions AND test to be above zero
- Multiplied predictions by 1.5



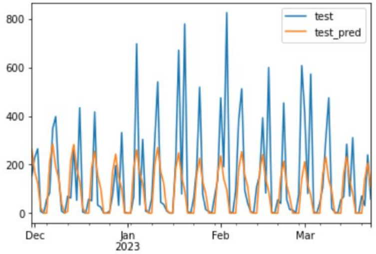
Midwest

Tuned Model

- Changepoint_prior_scale = 0.02
- yearly_seasonality = False
- .add_seasonality(name='weekly', period=7, fourier_order=1)
- .add_seasonality(name='monthly', period=30.5, fourier_order=1)

Test metrics

- Clipped test predictions AND test to be above zero
- Multiplied predictions by 2



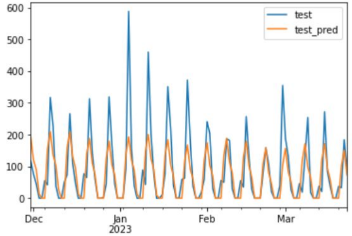
South

Tuned Model

- Changepoint_prior_scale = 0.02
- yearly_seasonality = False
- .add_seasonality(name='weekly', period=7, fourier_order=1)
- .add_seasonality(name='monthly', period=30.5, fourier_order=1)

Test metrics

- Clipped test predictions AND test to be above zero
- Multiplied predictions by 2



West

Tuned Model

- Changepoint_prior_scale = 0.02
- yearly_seasonality = False
- .add_seasonality(name='weekly', period=7, fourier_order=1)
- .add_seasonality(name='monthly', period=30.5, fourier_order=1)

Test metrics

- Clipped test predictions AND test to be above zero
- Multiplied predictions by 1.5

Univariate Model Key Findings

- Best models by region
  - East – Prophet
  - Midwest – Prophet
  - South – SARIMA
  - West – SARIMA

- Because there is a seasonality component to our data, the Prophet and SARIMA models performed better

- We found that setting yearly_seasonality = False worked better for our univariate prophets
  - Lessens the effect of spikes in January 2021 and 2022

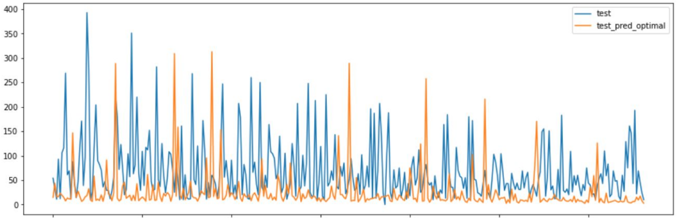
Multivariate Models

Multivariate Prophet

	MAE	MASE
Alissa	49.972	0.818
Emily	58.586	0.023
Nishi	77.815	0.0215
Ryan	35.980	0.589

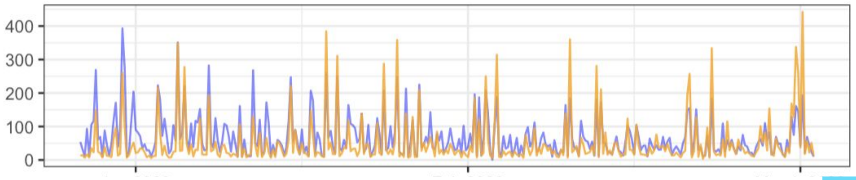
BEST MASE

- Missing vaccine data:
  - Extreme imputations so model can learn
- Tuning the model
  - yearly_seasonality = True, weekly_seasonality = True
  - changepoint_prior_scale=0.05, seasonality_prior_scale=10.0, n_changepoints=3
  - Added holidays (US)
  - Added regressors



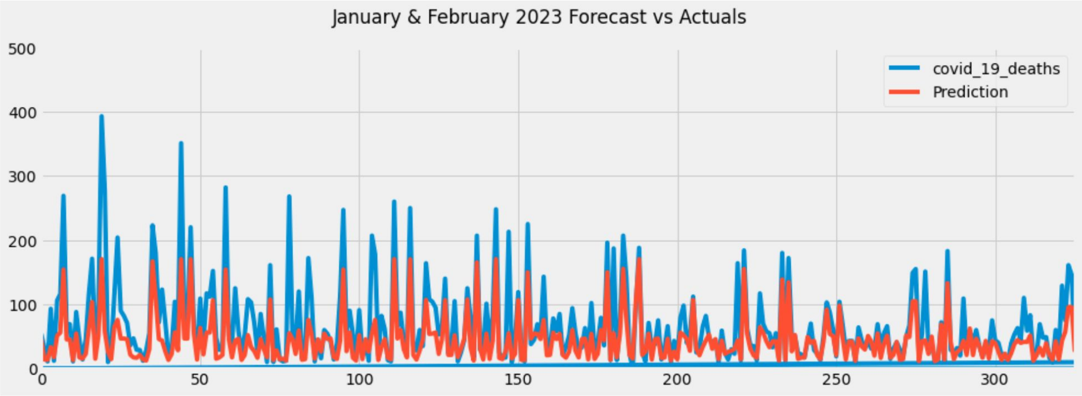
BEST MAE

- Missing vaccine data:
  - Extreme imputations so model can learn
- Tuning the model
  - yearly_seasonality = True
  - weekly_seasonality = True
  - n_changepoints=3
  - Added holidays (US)
  - Added regressors



XGBoost

	MAE	MASE
Alissa	16.199	0.00687
Emily	17.633	0.00748
Nishi	18.736	0.00794
Ryan	16.838	0.00745

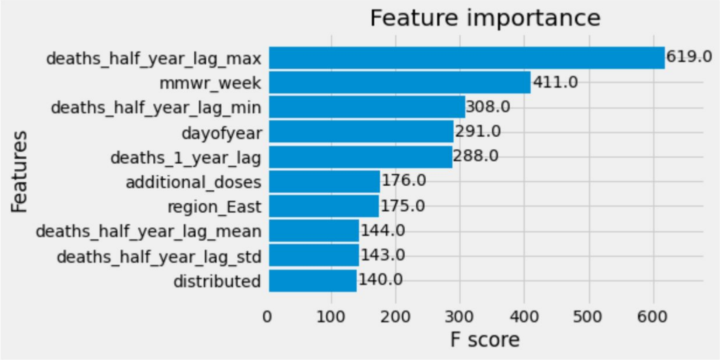


- Tuning Steps
  - Gridsearch with early stopping
  - 5 fold cross validation
- Baseline Model Performance
  - MAE: 26.132
  - MASE: 0.427
- Tuned Model
  - base_score = 0.5
  - learning_rate = 0.01
  - n_estimators = 100
  - early_stopping_rounds = 10

Multivariate Model Key Findings

- Prophet:
  - Yearly and Weekly Seasonality = TRUE
  - Holidays and regressors helped

- XGBoost was the best multivariate model
  - Most important features:
    - mmwr_week
    - lag / date features (half year, one year)
    - additional doses
    - distributed doses



Analysis & Conclusions

- Univariate Best Model
  - SARIMA (west)
- Multivariate Best Model
  - XGBoost
- Overall Best Performing Model
  - XGBoost
- Conclusions:
  - January 2021 and 2022 had spikes that led to overfitting and impacted our predictions for January 2023
  - Different regions in the US have varying results when model building
  - XGBoost can be helpful in predicting COVID-19 deaths, especially in regards to vaccination data and missing values

Model	**MASE**
ARIMA (west)	0.449
SARIMA (west)	0.373
Univariate Prophet (midwest)	0.436
Multivariate Prophet	0.022
XGBoost	0.00687

References

https://www.who.int/europe/emergencies/situations/covid-19

https://www.cdc.gov/museum/timeline/covid19.html

https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems

https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Week-Ending-D/r8kw-7aab/about_data

https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc/about_data

Thank You