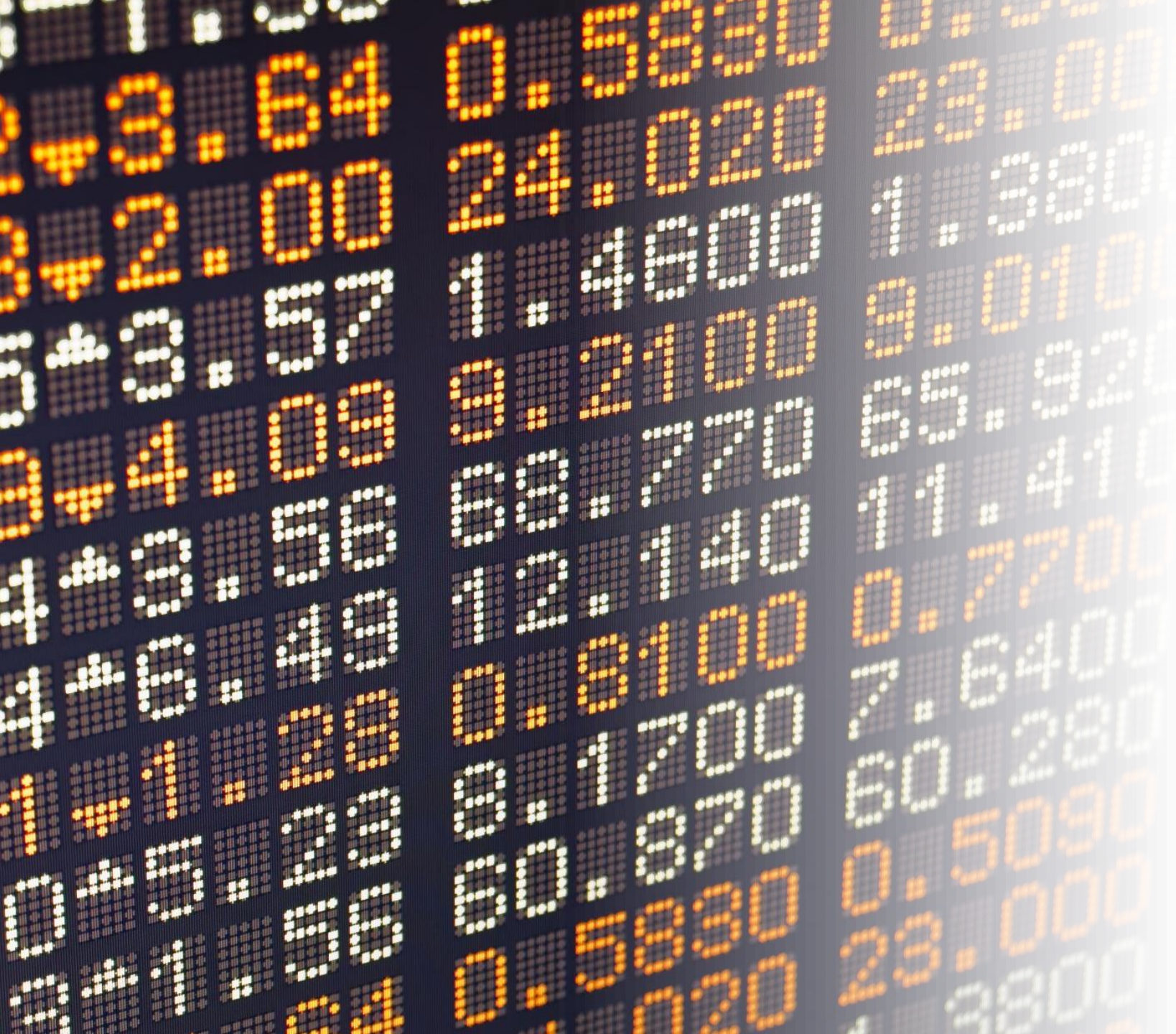


Forecasting

WHO Flu Data

Adnan Muhammed

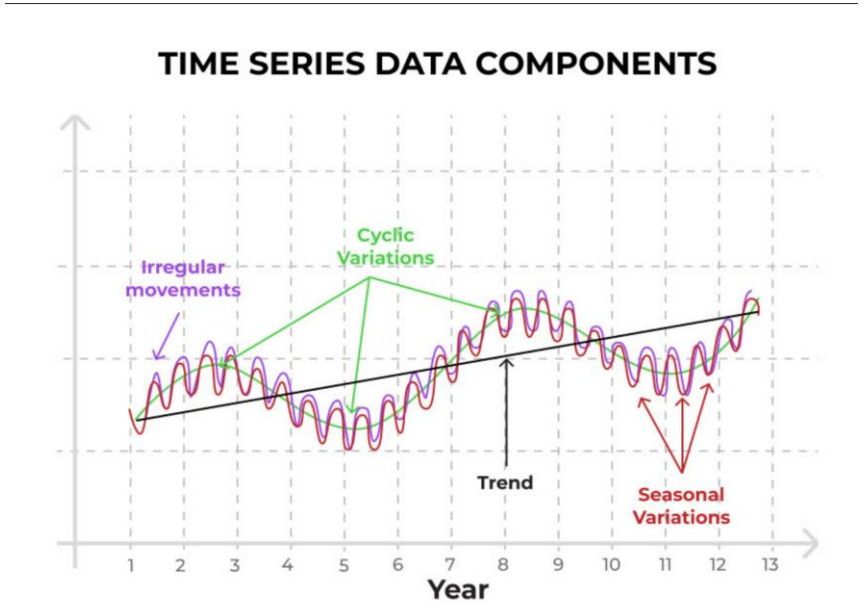
Agenda

- Introduction to Forecasting
- Forecasting Methods
- Industry Introduction and Use Cases
- Methods
- Code Walkthrough
- Challenges
- Conclusions

Intro to Forecasting

Making scientific predictions for future outcomes based on historical time data

- Building models through historical analysis and using them to drive future strategic decision-making
- Future outcome completely unavailable, must use evidence-based priors
- Applications include: Stock Price forecasting, Weather forecasting, Healthcare predictions, Business Planning



Forecasting Methods

MANIFOLD MARKETS

METACULUS
mapping the future

- **Online forecasting aggregation platforms:**
  - *Manifold Markets*
  - *Metaculus*
  - *Others*
- **Common models for forecasting:**
  - *Tree-based algorithms (Decision Trees, XGBoost)*
  - *Linear Regression*
  - *ARIMA*
  - *Exponential Smoothing Model*
  - *Neural Networks (LSTM, MLP, RNN, CNN)*
- **Univariate versus multivariate forecasting**
  - *Univariate Time Series Forecasting – only using previous values of time series*
  - *Multivariate Time Series Forecasting – using predictors other than the series (a.k.a exogenous variables)*

Industry Introduction and Use Cases

Forecasting in Healthcare

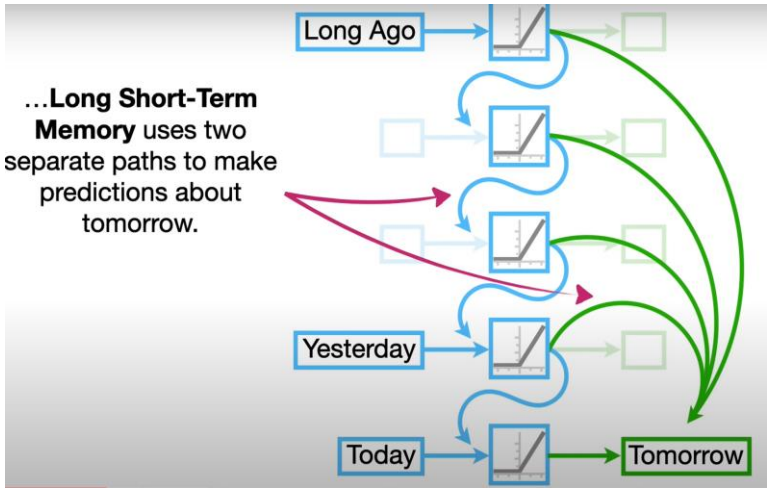
- Lots of publicly available data
- Predicting number of COVID deaths, flu cases, hospital beds, etc.
- Used for seasonal illnesses
- Monitoring and predicting disease spread
- Informing decisions about resource allocation (vaccines, traveling nurses, etc.)

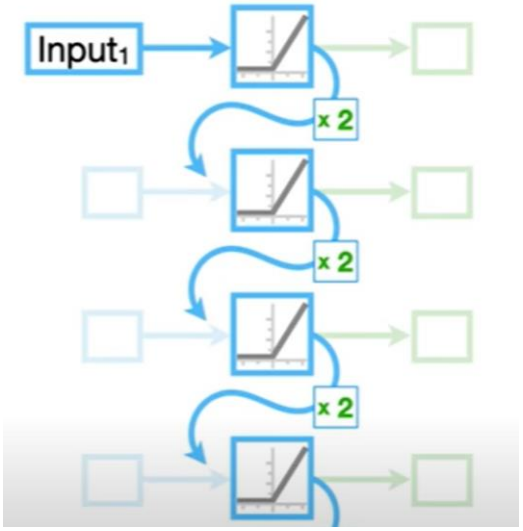
Models in Healthcare

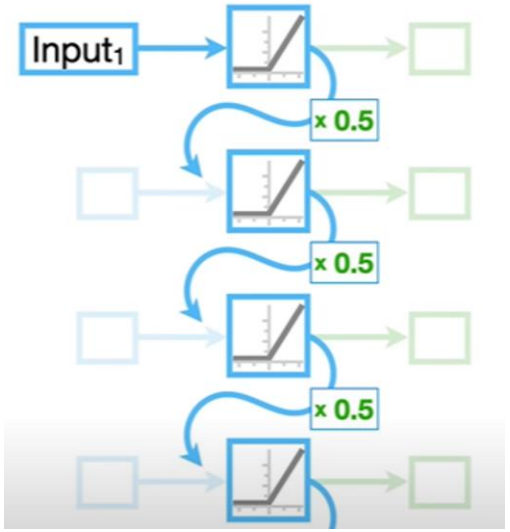
- ARIMA / SARIMA
- LSTM
- Linear Regression
- Multilayer perceptron
- Vector autoregression

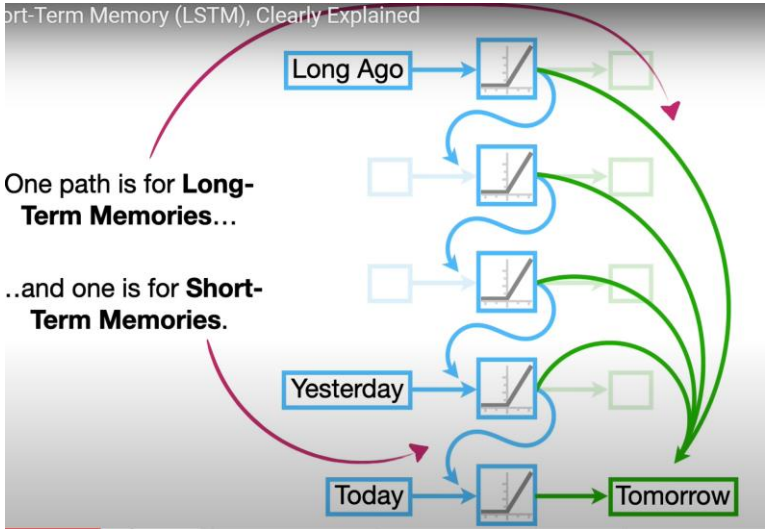
Forecasting Methods: Neural Networks

- RNN: using the same feedback loop connection, so the gradient can explode or vanish. It uses different paths for long and short-term memories.
- LSTM: a type of RNN, which is designed to avoid exploding or vanishing



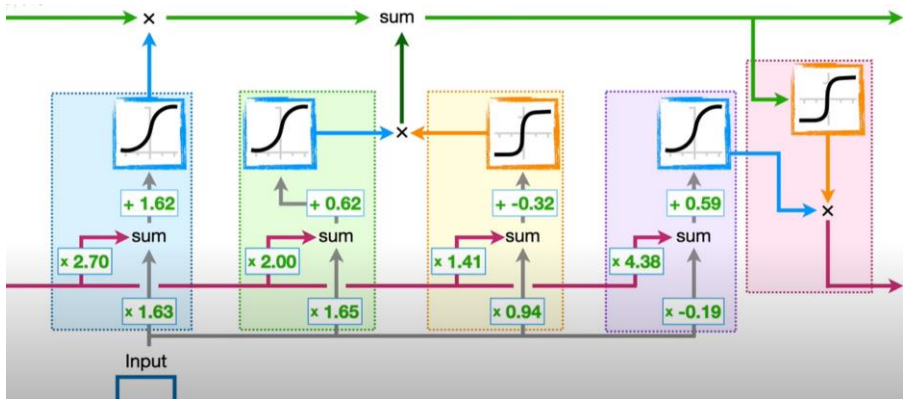


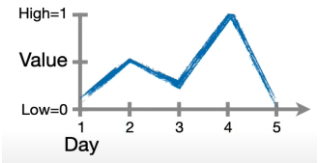


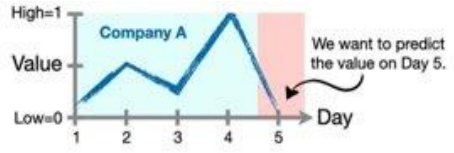


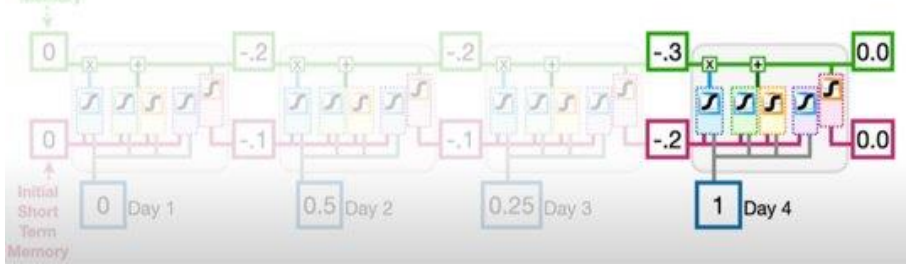
Forecasting Methods: LSTM

How the LSTM unit works



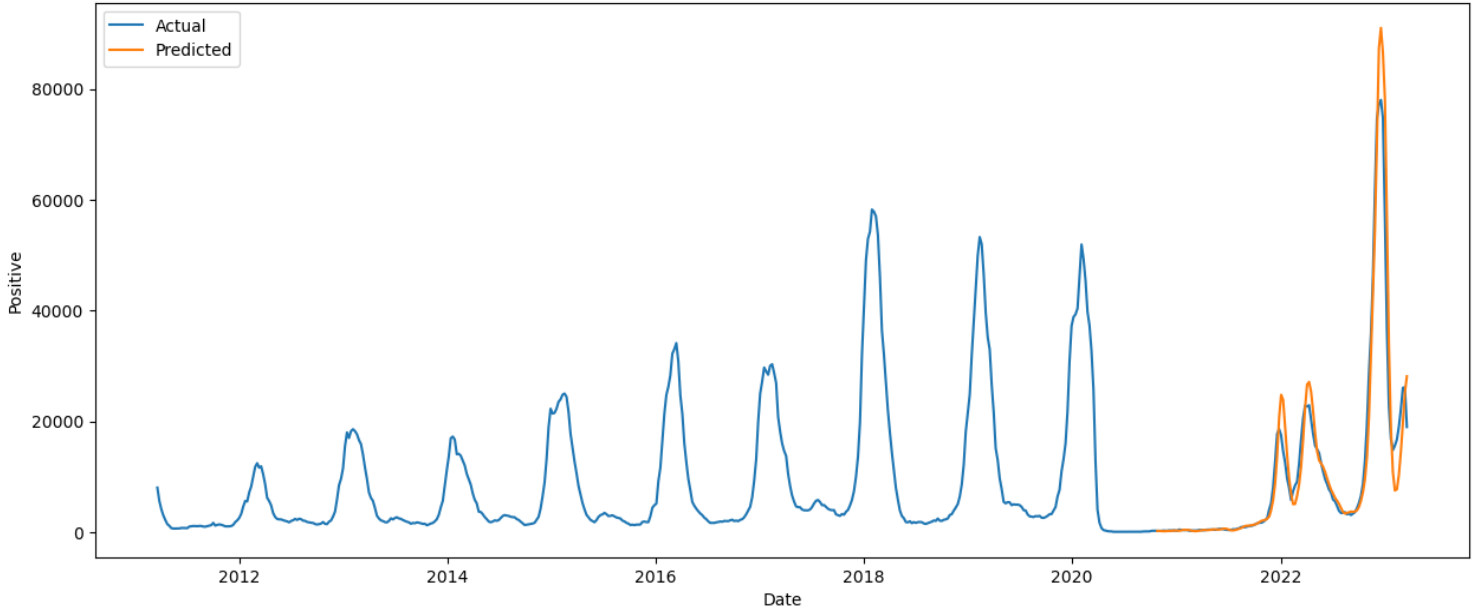






Example: Forecasting stock price at Day5

- Our work using LSTM for influenza forecasting
- Good forecasting
- Complicated and time consuming (with the help of GPT)



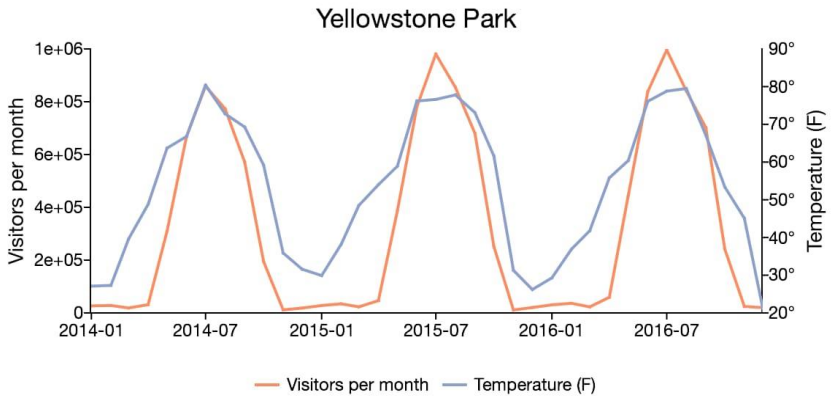
Forecasting Methods: ARIMA

Autoregressive Integrated Moving Average Model

- a form of regression analysis that gauges the strength of one dependent variable relative to other changing variables.
- ARIMA models use differencing to convert a non-stationary time series into a stationary one, and then predict future values from historical data.

Variations:

- SARIMA
  - Seasonal ARIMA
- SARIMAX
  - SARIMA using exogenous variable(s)



ARIMA

Parameters:

- Autoregression (p) - a model that shows a changing variable that regresses on its own lagged, or prior, values.
- Integrated (d) - the differencing of raw observations to allow the time series to become stationary (i.e., data values are replaced by the difference between the data values and the previous values).
- Moving average (q) - dependency between an observation and a residual error from a moving average model applied to lagged observations.

Pros and Cons:

Pros

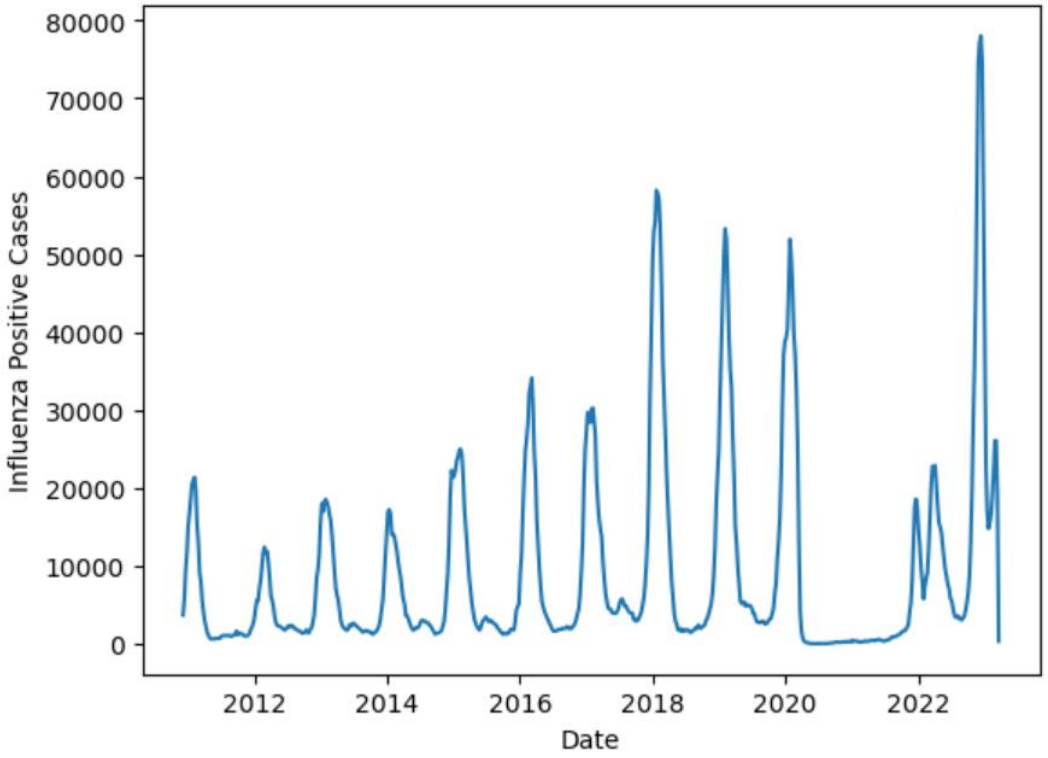
- Good for short-term forecasting
- Only needs historical data
- Models non-stationary data
- Easy to understand and interpret

Cons

- Not built for long-term forecasting
- Poor at predicting turning points
- Computationally expensive
- Parameters are subjective

Model Training: Dataset

- Historical data of flu infections worldwide from the WHO
- Size: 109083 rows, 23 columns
- Timespan: 2010-12 to 2023-03



World Health Organization

INFLUENZA LABORATORY SURVEILLANCE INFORMATION
Influenza virus detections reported to FluNet

GISRS GLOBAL INFLUENZA SURVEILLANCE & RESPONSE SYSTEM

Date last refreshed (UTC)
3/29/2023 7:01:28 PM

Country, area or territory: All
WHO region: All
Influenza transmission zone: All
Surveillance site type*: All

To export the data, use the right click on your mouse and select 'Open link in a new tab' or 'Save link as'. If the page freezes, please select less records.

Data dictionary

Week start date: 9/27/2021 – 3/20/2023

17027 records selected

									Number of influenza A viruses detected by subtype					Number of influenza B viruses detected by lineage						
WHO region	Influenza transmission zone	Country, area or territory	Surveillance site type	Year	Week	Start date	Specimens received	Specimens processed	A (H1)	A (H1N1) pdm09	A (H3)	A (H5)	A (not subtyped)	A (Total)	B (Yamagata lineage)	B (Victoria lineage)	B (not determined)	B (Total)	Influenza positive specimens	Influenza negative specimens	ILI activity
AFR	Eastern Africa	Madagascar	Sentinel	2023	12	2023-03-20	42	42	0	0	0	0	3	3	0	2	0	2	5	37	Sporadic
AFR	Eastern Africa	Zambia	Type not defined	2023	12	2023-03-20	0	0	0	0	0	0	0	0	0	0	0	0	0		No Report
AFR	Middle Africa	Cameroon	Sentinel	2023	12	2023-03-20	27	27	0	0	0	0	0	0	0	1	0	1	1	26	No Report
AFR	Middle Africa	Democratic Republic of the Congo	Sentinel	2023	12	2023-03-20	26	26	0	0	0	0	0	0	0	0	0	0	0	26	No Report
AFR	Western Africa	Burkina Faso	Sentinel	2023	12	2023-03-20	38	38	6	0	0	0	6	6	0	0	0	0	6	32	No Report
AFR	Western Africa	Côte d'Ivoire	Type not defined	2023	12	2023-03-20	29	29	2	0	0	0	0	2	0	0	0	0	2	27	Sporadic
AFR	Western Africa	Senegal	Sentinel	2023	12	2023-03-20	17	17	0	0	0	0	0	0	0	0	0	0	0	17	No Report
AFR	Western Africa	Togo	Sentinel	2023	12	2023-03-20	7	7	0	0	0	0	0	0	0	0	0	0	0	7	No Report
AMR	Central America and Caribbean	Anguilla	Sentinel	2023	12	2023-03-20		0						0				0	0	0	No Report
AMR	Central America and Caribbean	Antigua and Barbuda	Sentinel	2023	12	2023-03-20		0						0				0	0	0	No Report
AMR	Central America and Caribbean	Aruba	Sentinel	2023	12	2023-03-20		0						0				0	0	0	No Report
AMR	Central America and	Bahamas	Sentinel	2023	12	2023-03-20		0						0				0	0	0	No Report

***Surveillance site type:**
- **Non-sentinel**: Data obtained from non-sentinel systems as indicated by the reporting country. Data reported in this category may include outbreak investigation, universal testing, testing at point of care or other systems apart from sentinel surveillance.
- **Sentinel**: Data obtained from sentinel surveillance as indicated by the reporting country. Sentinel surveillance systems collect high-quality data in a timely manner systematically and routinely from sentinel surveillance sites representatives of the population under surveillance.
- **Type not defined**: Source of data not indicated by the reporting country neither as sentinel nor as non-sentinel surveillance. These data may include sentinel or non-sentinel surveillance sources or both.

Calendar type: ISO 8601

Data source: FluNet (https://www.who.int/tools/flunet)

Back to Top

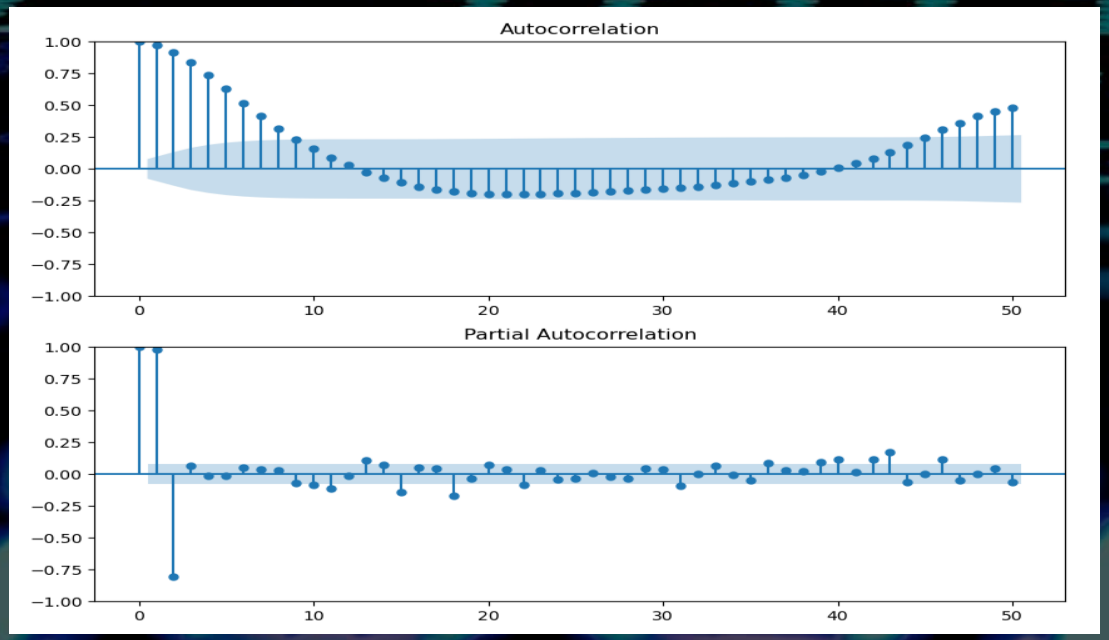
```
In [221]: df = df.drop(columns = ['ISO_WEEK', 'SPEC_RECEIVED_NB', 'SPEC_PROCESSED_NB','AH1',
                                  'AH5', 'INF_A', 'ANOTSUBTYPED', 'BVIC', 'BYAM', 'BNOTDETERM
                                  'INF_NEGATIVE', 'ILI_ACTIVITY', 'WHOREGION', 'ITZ', 'COUNTR
                                  'COUNTRY_CODE','ORIGIN_SOURCE','ISO_YEAR']) # Dropping all

In [222]: df.info()
```

Code Walkthrough

```
Augmneted Dickey_fuller Statistic: -7.814637
p-value: 0.000000
critical values at different levels:
        1%: -3.441
        5%: -2.866
        10%: -2.569
```

Date	Positive
2010-12-05 18:30:00+00:00	3699.0
2010-12-12 18:30:00+00:00	5755.0
2010-12-19 18:30:00+00:00	9384.0
2010-12-26 18:30:00+00:00	11594.0
2011-01-02 18:30:00+00:00	15185.0



ARIMA

```
model = ARIMA(df['Positive'], order=(1,1,1))
results = model.fit()
forecast = results.predict(start='2018-01-21 18:30:00+00:00', end='2023-03-19 18:30:00
# plt.plot(df['Positive'])
```

ARIMA Seasonal Ordered

```
model = ARIMA(df['Positive'], order=(1,1,1), seasonal_order=(1,1,1,52))
results = model.fit()
forecast = results.predict(start='2018-01-21 18:30:00+00:00', end='2023-03-19 18:30:00+00:00
# plt.plot(df['Positive'])
```

A Basic Sarimax model is now run to see if it fits better.

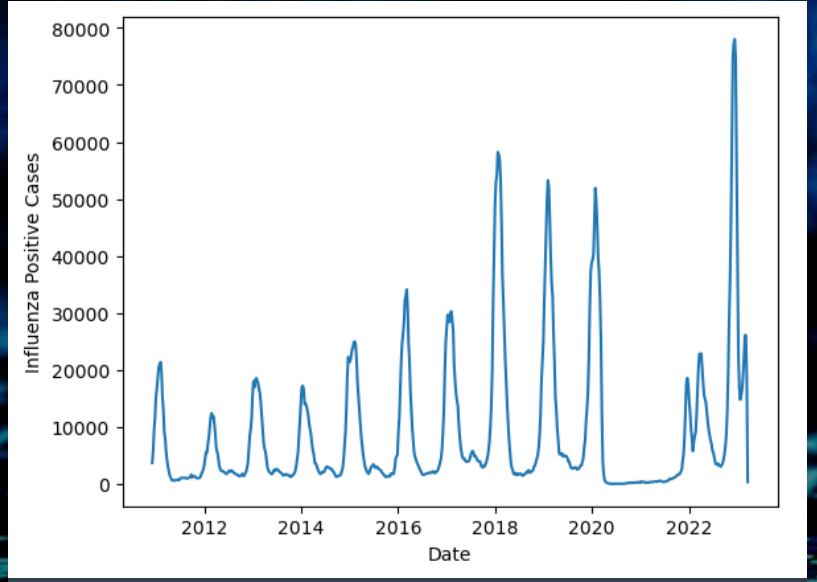
```
# Sarima Model
model = sm.tsa.statespace.SARIMAX(df, order=(1,1,1), seasonal_order=(1,1,1,52))
```

A Model is Tuned using ranges 1 to 5 for the order.

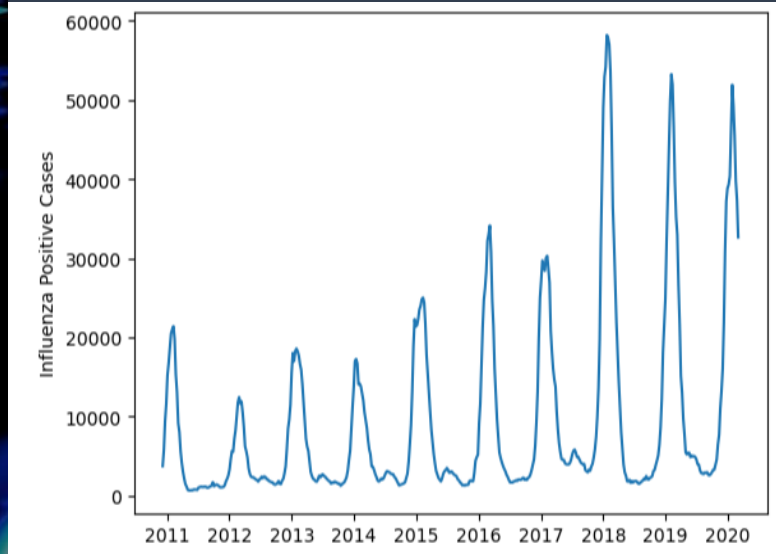
```
# Model Tuning using AutoArima
tunedmodel = pm.auto_arima(df, start_p=1, start_q=1, max_p=5, max_q=5, m=52, seasonal=True
                           trace=True, error_action='ignore', suppress_warnings=True, stepwise=
```

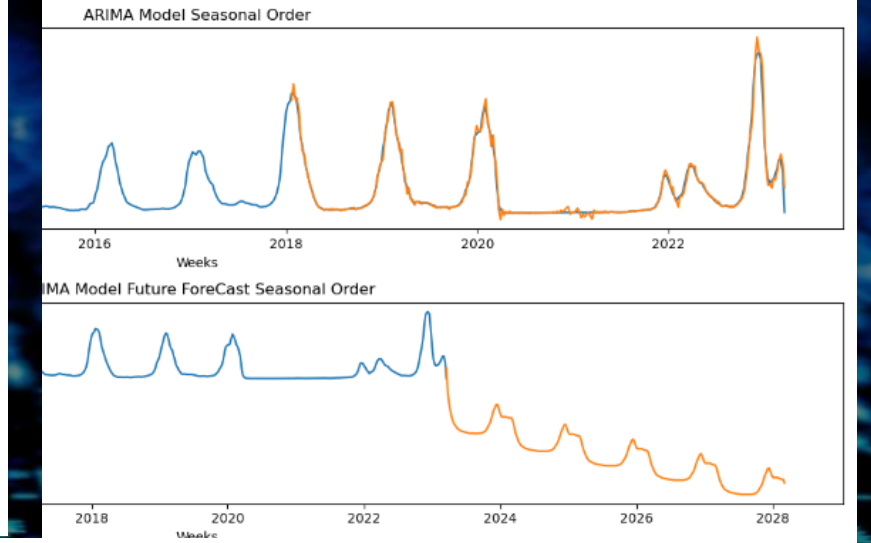
```
Performing stepwise search to minimize aic
 ARIMA(1,1,1)(1,0,1)[52] intercept   : AIC=11222.623, Time=1.88 sec
 ARIMA(0,1,0)(0,0,0)[52] intercept   : AIC=11985.915, Time=0.04 sec
 ARIMA(1,1,0)(1,0,0)[52] intercept   : AIC=11267.792, Time=1.07 sec
 ARIMA(0,1,1)(0,0,1)[52] intercept   : AIC=11476.258, Time=2.69 sec
 ARIMA(0,1,0)(0,0,0)[52]             : AIC=11983.918, Time=0.04 sec
```

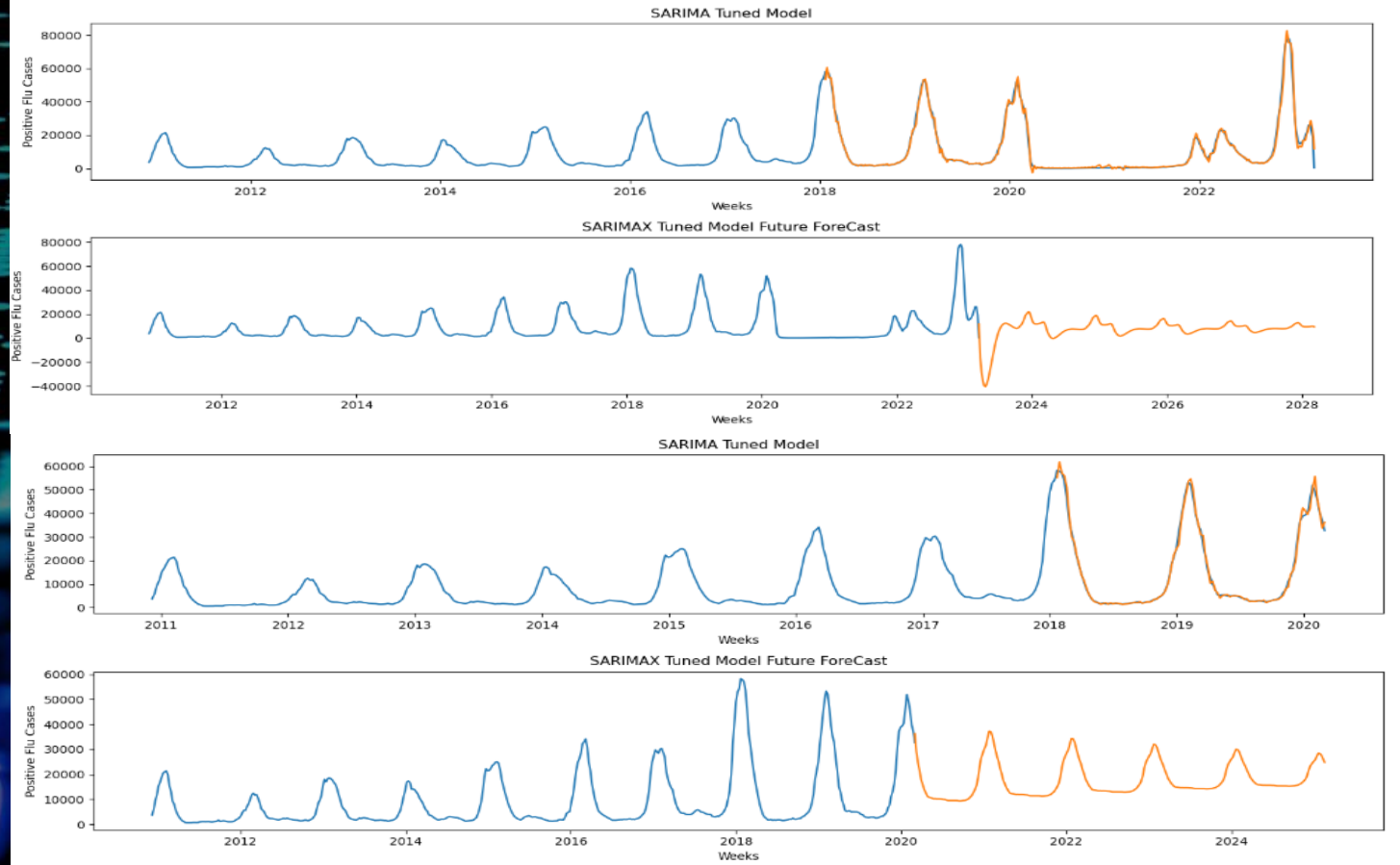
```
# fit a SARIMAX model using the optimal order parameters
model = sm.tsa.statespace.SARIMAX(df, order=tunedmodel.order, seasonal_order=tuned
results = model.fit()
forecast = results.predict(start='2018-01-21 18:30:00+00:00', end='2023-03-19 18:3
```



Graph Walkthrough







Forecasting Metrics

- Mean Squared Error (MSE)
  - penalizes large errors or outliers more compared to the small errors
- Root Mean Squared Error (RMSE)
  - RMSE value is in the same unit as the forecasted value
  - easier to understand compared to MSE
- Mean Absolute Percent Error (MAPE)
  - measures the average of forecast errors in percentages
  - works best with data without zeros and extreme values
- Root Mean Squared Log Error (RMSLE)
  - not scale-dependent and is useful across a range of scales
  - It is not affected by large outliers.
  - It considers only the relative error between the actual value and the predicted value.

Metrics used to evaluate the final model

Mean Squared Error (MSE)	2564043.054
Root Mean Squared Error (RMSE)	1601.262956
Mean Absolute Percentage Error (MAPE)	7.837775809
Root Mean Squared Log Error (RMSLE)	0.101783719

Other metrics used in forecasting

- Huber Loss
- Log Cosh Loss
- Quantile Loss

Challenges and Conclusions

Challenges

- Publicly available data messy or limited
- Hyperparameter tuning
- Disruption in seasonal patterns from COVID pandemic
- Different metrics for forecasting models (no precision, recall, F1)
- Google Collab problems
- LSTM failed lots of times, very slow
- Choosing between ARIMA and SARIMA
- Controlling Dataset

Conclusions

- Forecasting is a powerful tool
- Time series models are unique
- MSE lower for SARIMA than ARIMA
- Dataset not necessarily accurate due to Covid Disruptions

Next Steps

- Improve Model using Exogenous variables.
- Learn to read ACF & PACF Plots better to tune orders.