Where's the next dengue cluster?

A dengue case prediction project by
Zengdi, Zhiming, Priscilla and Daphne

Overview

01 Background

02 Problem Statement

03 Data Science Approach

04 Datasets & EDA

05 Modelling

06 Conclusions

Why is dengue dangerous?

- Dengue fever is a disease prevalent in the tropics. It is caused by the dengue virus, which in turn is transmitted to humans via the bite of an infected Aedes mosquito

- WHO ranked dengue as the "most important mosquito-borne viral disease in the world" with an estimated **40 per cent of the world population at risk, or 100–400 million infections each year.**

- Estimated economic impact of dengue to Singapore from 2010-2020: **1-2 Billion USD or 7,000 - 21,000 DALYs.** Singapore had experienced **two-year dengue outbreaks** in 2013 and 2014, 2015 and 2016

- Common symptoms are high fever, headache, body aches, nausea and rash. **Severe cases, dengue can be fatal.** There is no specific treatment for dengue. Prevention and vector control is relied on to reduce incidence of dengue.

NEA's efforts to stamp out dengue

- **National Dengue Prevention Campaign**
  - Mozzie Wipeout B-L-O-C-K steps
  - S-A-W protective steps
- Work with Town Councils to coordinate **chemical treatment**, such as fogging, misting and larviciding
- Conduct an average of **800,000 inspections per year**. Premise owners subjected to $5K fine or up to 3 months jail.
- **Deployed 70,000 Gravitraps** designed to attract and trap female *Aedes* adult mosquitoes that are looking for sites to lay their eggs in
- Pilot Wolbachia-carrying Aedes Mosquito suppression project

Cost Constraint

- **NEA KPI: <300 local dengue fever cases per 100,000 population:**
- **NEA annual dengue prevention budget: SGD 60 million**

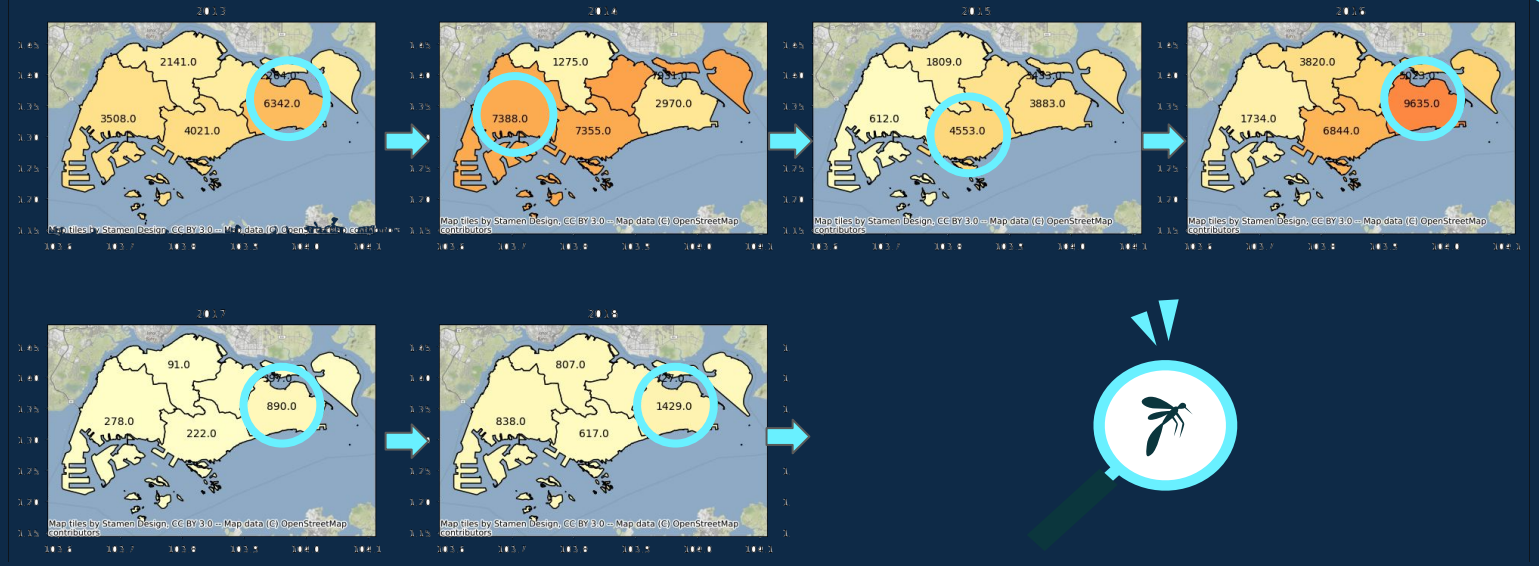
SN	Description	Cost
1	Community Outreach (National Dengue Prevention Campaign)	$12 million
2	**Operating Expense - Inspections**	**$30 million**
3	**Operating Expense - Chemical Treatment**	**$8 million**
4	Wolbachia Project + Research	$3 million
5	**Trapping + surveillance**	**$7 million**
	Total Cost:	$60 million

For each 55 planning areas:

- **4 inspections / man-day**
- **Monthly fogging**
- **75 gravitraps locations**

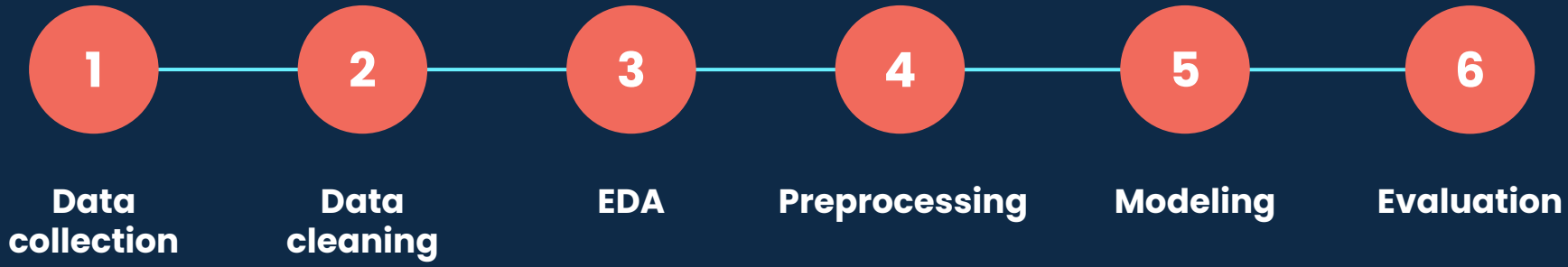
Problem Statement

Annual No. of Dengue Cases from 2013 onwards



- Forecast the next dengue cluster region out of the five: *West, North, North-East, East and Central,* and guide NEA in allocating operational resources.

Data Science Approach



Datasets

Weather data	Population density data	Dengue data	Google data
Data collected by Meteorological Services Singapore	Data collected via Onemap.sg API services	Data collected from NEA's website	Google search data
Rainfall, Temperature, Wind data 2013-2018 Relative Humidity 2016-2018	No. of res/hlds by housing type and planning area 2010, 2015, 2018	Reported cases by clusters (locations) 2013-2018	Search terms containing 'dengue', 'mosquito', 'insect repellent' and so on 2013-2018

Data cleaning

Treatment

- Numerical values are aggregated based on year and week number.
- Locational values are aggregated based on 5 regions.

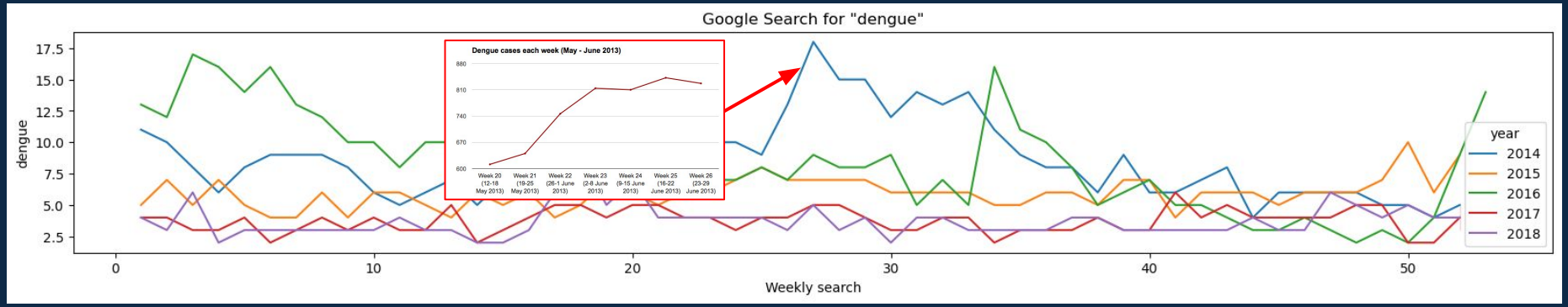
Missing values

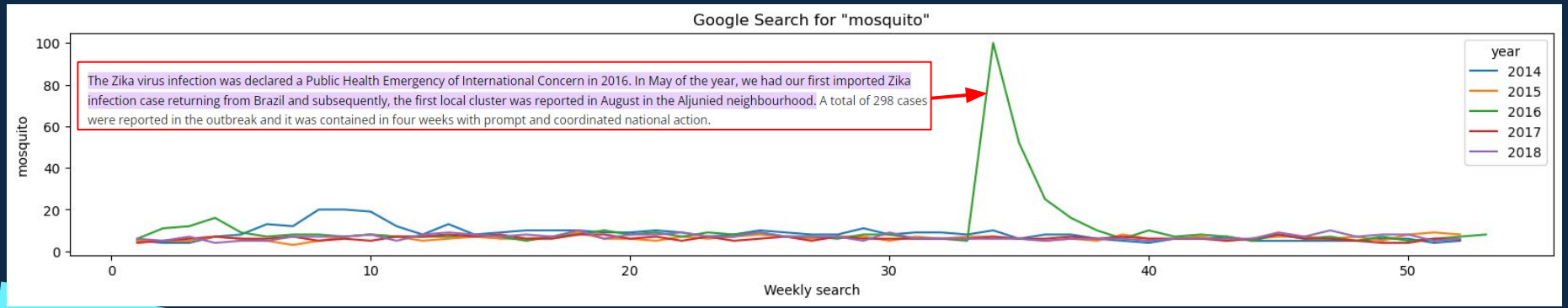
- Backfill: missing years in weather data
- Imputing the average of the nearest values: missing dengue data
- Fill with 0: NAH in Google search data

Outliers

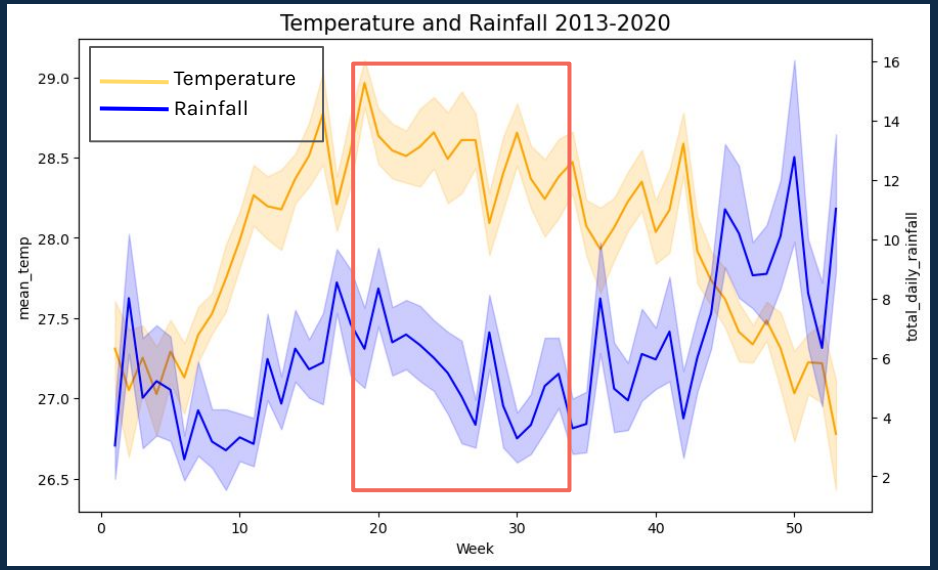
- Outliers (especially for certain regions) are not removed.

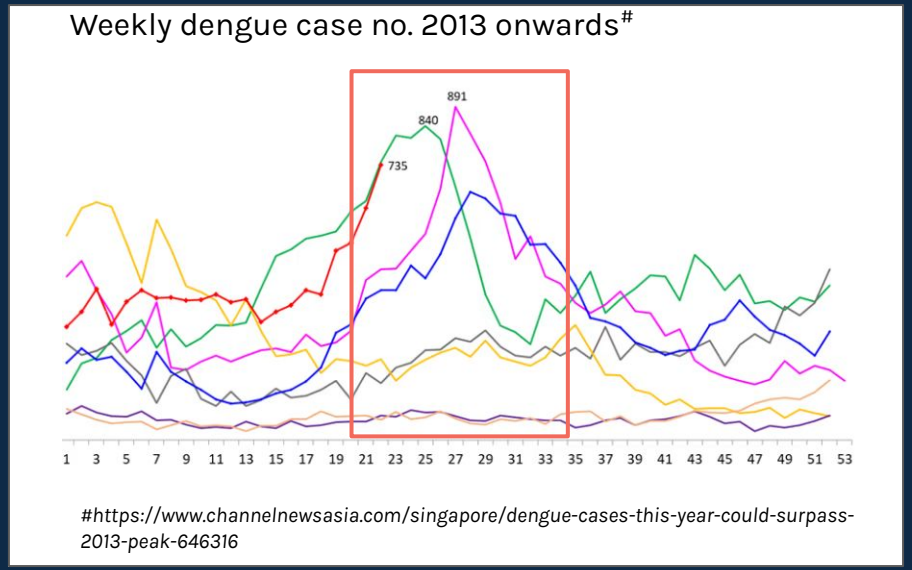
EDA – google search trend





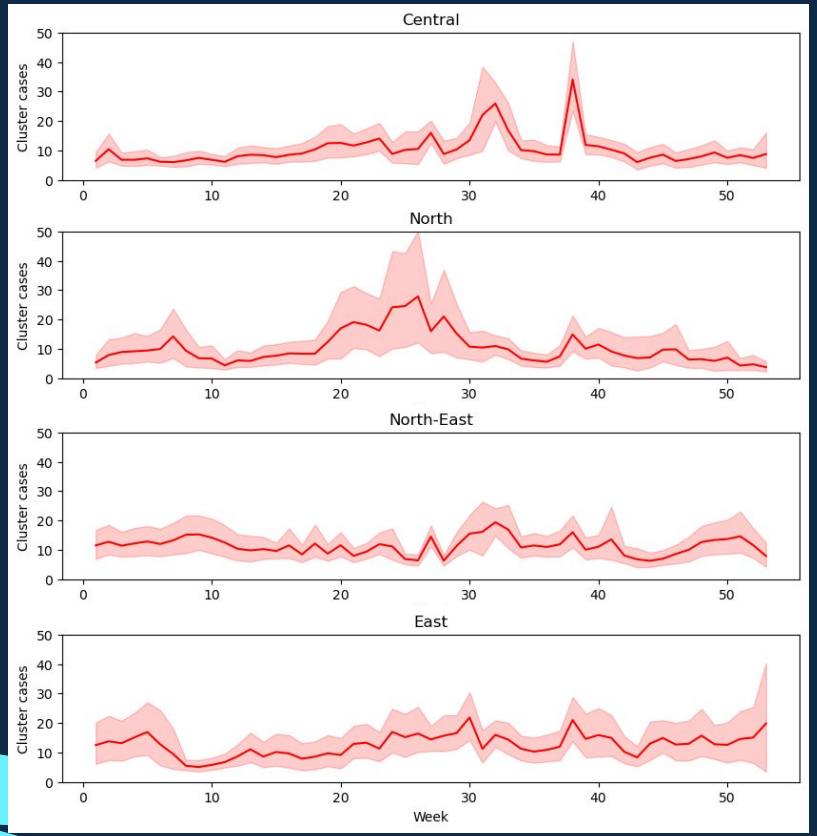
EDA – weather data

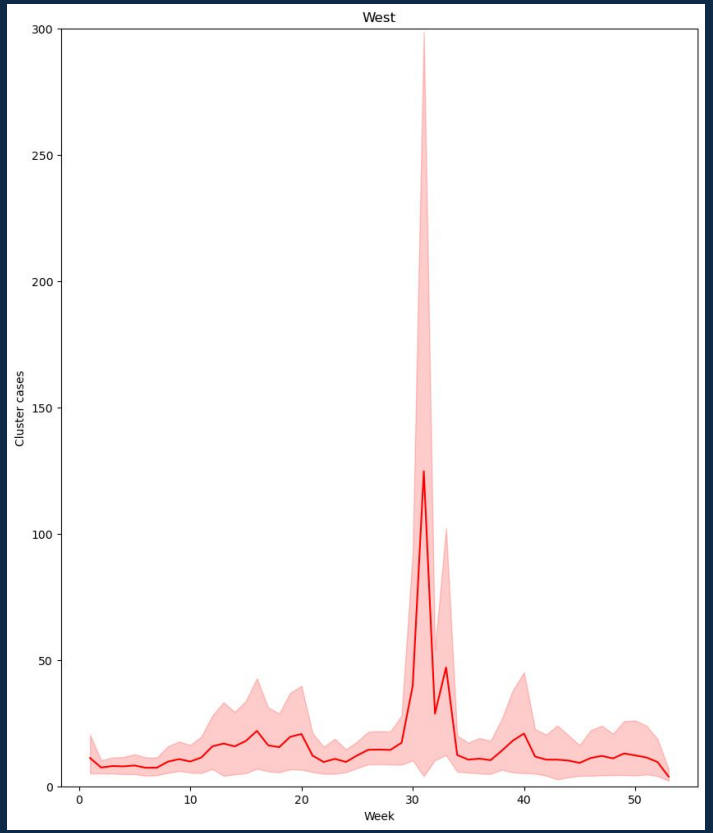




#https://www.channelnewsasia.com/singapore/dengue-cases-this-year-could-surpass-2013-peak-646316

EDA - dengue clusters by regions





EDA - Singapore Dengue Hotspot Map (2013-2018)

Before Modelling

Preprocessing

- Divide dataset into 5 geographical regions
- Shifting of data points
- Min Max Scaling

Feature Selection

- Backward feature selection (Linear Regression)
- Pruning (RF)
- ACF/ PACF (LSTM)

Model Selection

Model	Pros	Cons
Random Forest	Flexible (less sensitive to outliers)	Less interpretable
Negative Binomial Regression	Can accommodate a higher degree of variability in the data	unreliable predictions beyond the range of the observed data.
Linear Regression	Interpretable	Independence assumption
Long Short Term Memory	Time series in nature, Better at handling long-term dependencies	High model complexity, May not work well with highly nonlinear data or data with a lot of noise

Model 1: LSTM

Step 1: Normalise data (Min Max Scaling)

Step 2: Reshape data into the correct shape

Step 3: Train-test split the reshaped data

Step 4: Create a model using Keras

Step 5: Train the model

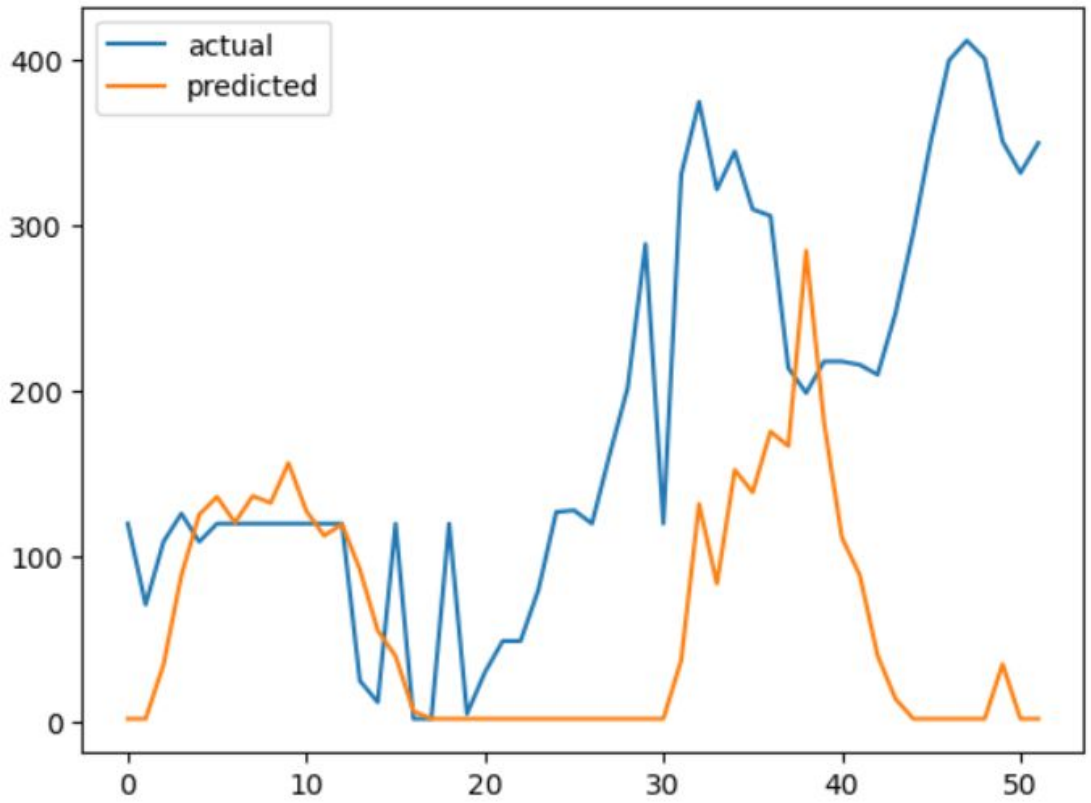
Step 6: Make inference from trained model

Step 7: Scale result of inference back to original

Step 8: Examine the results

Model 1: LSTM - reshape data

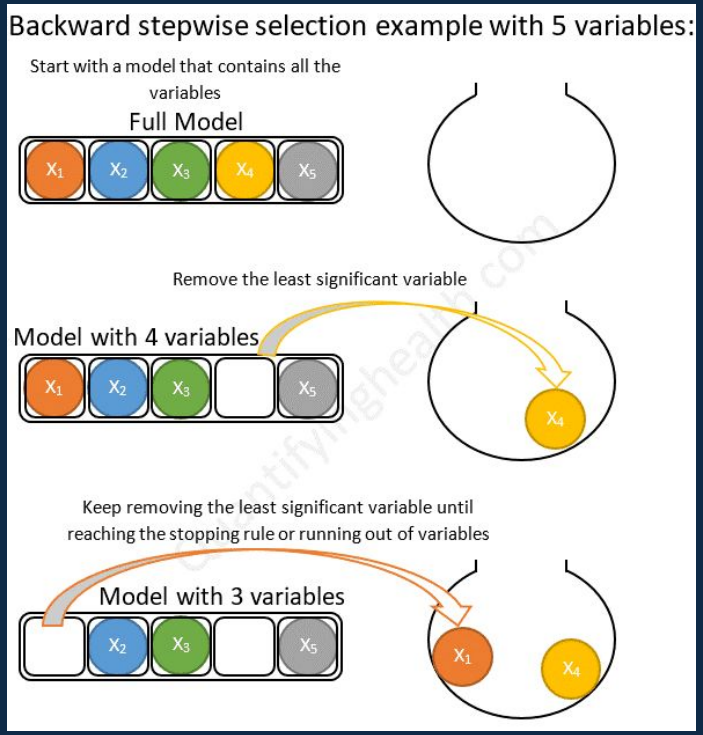
Model 1: LSTM – model results (West)

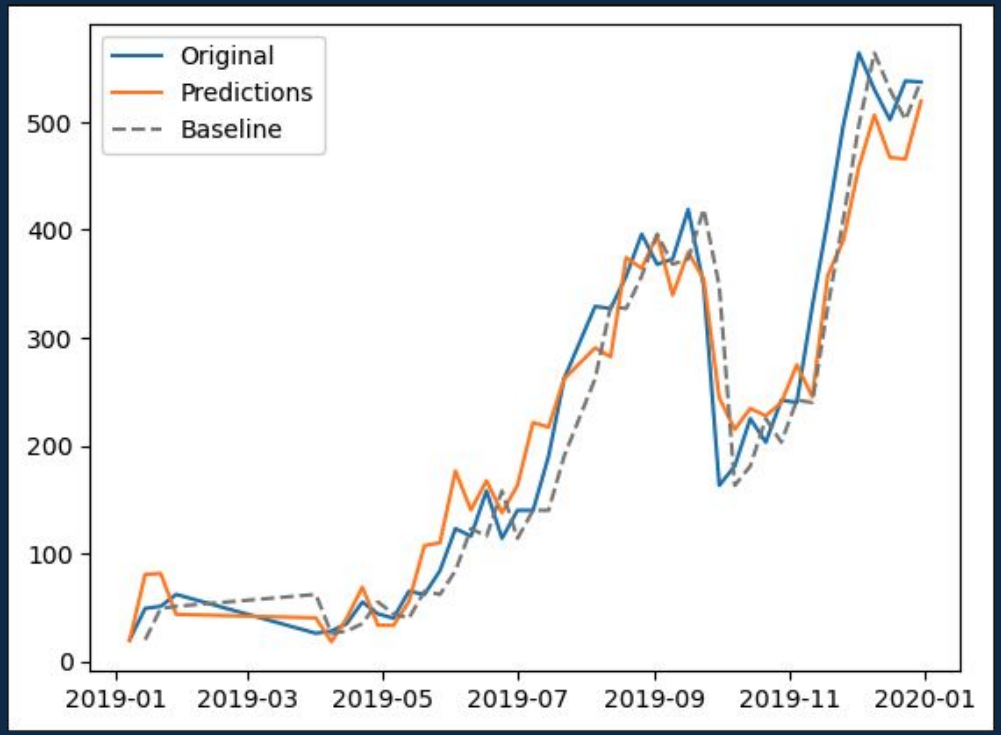


Model for West:
Train:
RMSE: 64
MAE: 35

Test:
RMSE: 181
MAE: 134

Model 2: LR - With backward stepwise selection (NE)

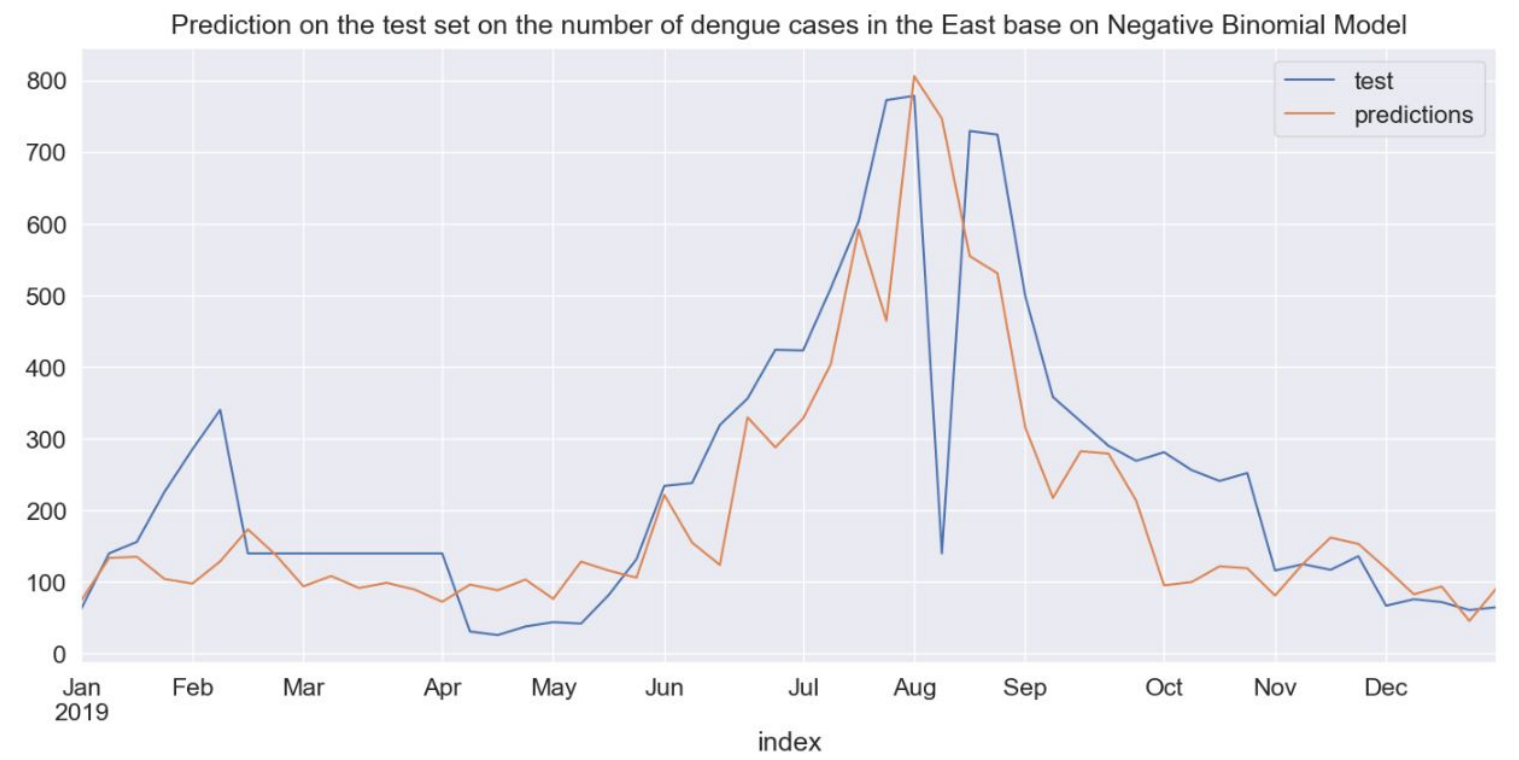




RMSE: 41.85 (Test)

MAE: 32.26 (Train), 32.26 (Test)

Model 3: Negative Binomial (NB) regression model(East)



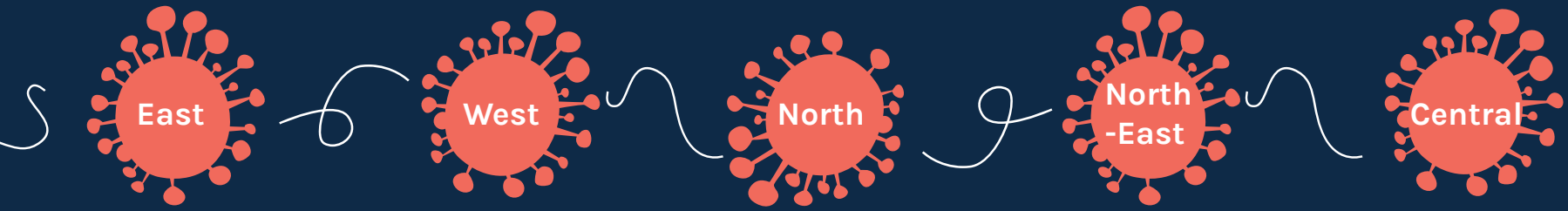
RMSE: 131.97(Test)

MAE: 236.14 (Train), 123.01 (Test)

Model Comparison

Region	Model	Train MAE	Test MAE
West	ForecasterAutoreg model	73.37	154.94
	Ridge Regression Model	75.42	102.63
	Persistence Model	**28.4**	**28.5**
North	**ForecasterAutoreg model**	**28.05**	**23.05**
	Negative Binomial Model	28.61	86.44
	Persistence Model	16.5	30.0
Northeast	ForecasterAutoreg model	70.10	152.25
	Negative Binomial Model	71.23	140.05
	Linear Regression with backward stepwise feature selection	**32.2**	**32.3**
East	ForecasterAutoreg model	90.64	157.86
	Negative Binomial Model	236	123
	OLS with backward stepwise feature selection	**51.0**	**51.0**
Central	ForecasterAutoreg model	105.97	192.98
	Negative Binomial Model	421.30	137.92
	Linear Regression with backward stepwise feature selection	**42.2**	**42.2**

Ensemble Models



OLS with backward stepwise feature selection	Persistence Model	ForecasterAutoreg model	Linear Regression with backward stepwise feature selection	Linear Regression with backward stepwise feature selection
Predicted cluster cases: **13,094**	Predicted cluster cases: **8,083**	Predicted cluster cases: **5,473**	Predicted cluster cases: **9,405**	Predicted cluster cases: **10,008**

Limitations

Data limitation:

- Unlike the nationwide data, dengue cluster numbers are registered in an accumulative fashion including both new and existing cases. The numbers are indicative only.

Insufficient data points:

- More data points that provide an entire sequential variables of the dengue events would likely improve the performance of a few models including LSTM.

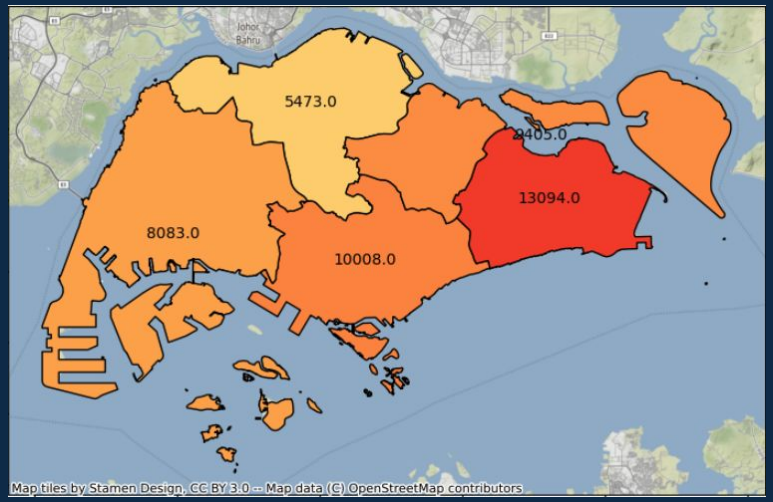
Uncertainty with the lag:

- Time series and regression models that take in the lag feature would significantly improve with a known lag.

Not all factors are included in the dataset.

Recap: Business Problem

- NEA KPI: <300 local dengue fever cases per 100,000 population:
- NEA annual dengue prevention budget: SGD 60 million



Impact of estimate reduction of 150 dengue cases	
Item	Value
Cost savings (medical expenses)	$800,000
Cost savings (lost productivity)	$1,000,000
Total Benefits	$1,800,000

Thank You!

Do you have any questions?