Department of Industrial Engineering

CHULA ENGINEERING
Foundation toward Innovation

Senior Project Presentation

CHULA ENGINEERING
Foundation toward Innovation

Predicting Dengue-Virus Disease Spread

Pasin Sirirat 5930351421

Agenda

- Introduction
- Exploratory Data Analysis
- Modeling
- Results
- Conclusion

Introduction

Dengue Fever

- A mosquito-borne disease caused by Dengue Virus, spread through the bites of infected *Aedes* species mosquitoes.
- The symptoms varies from mild fever to bleeding and death.

CHULA ENGINEERING
Foundation toward Innovation

How many people are affected?

Objectives

- To construct the prediction model to forecast the number of dengue case occurrences in the cities of San Juan and Iquitos.

- To compare prediction techniques between the Time-Series Forecasting and the Machine Learning methods.

CHULA ENGINEERING
Foundation toward Innovation

DRIVENDATA

Data

- The data comes from the DrivenData online data science competition.
- The numbers of cases in both cities came from the Center for Disease Control (CDC).
- The features used to build models are 20 climate variables which came from four sources:
  - Normalized Difference Vegetation Index (NDVI)
  - Precipitation Estimation from Remotely Sensed Information using an Artificial Neural Network (PERSIANN)
  - NOAA's Reanalysis Method
  - Global Historical Climatology Network (GHCN) Weather Stations
- The data in San Juan comprises of 936 records, whereas the data in Iquitos comprises of 520 records, both are in weekly format.

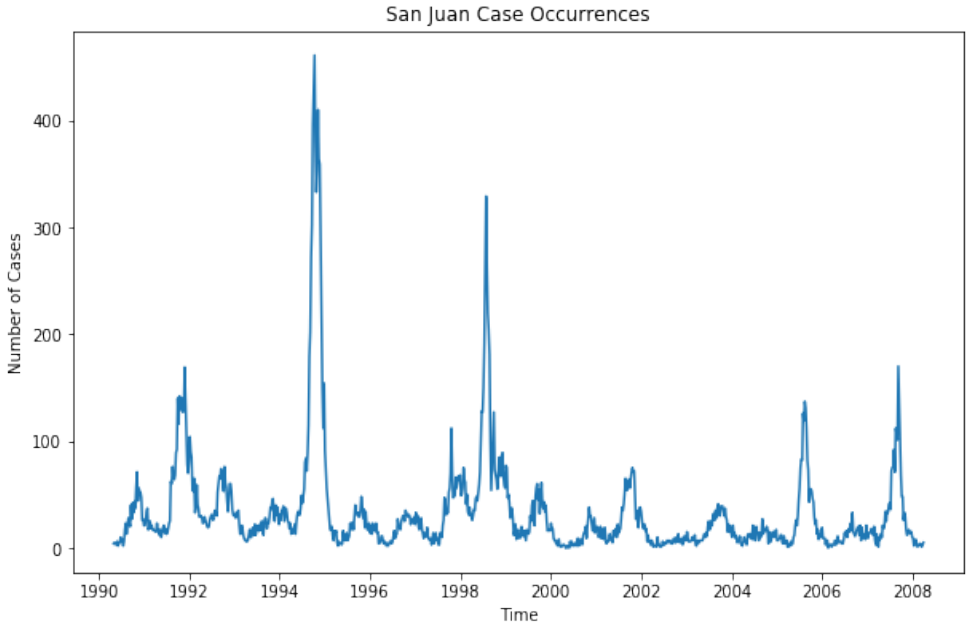
Performance Evaluation

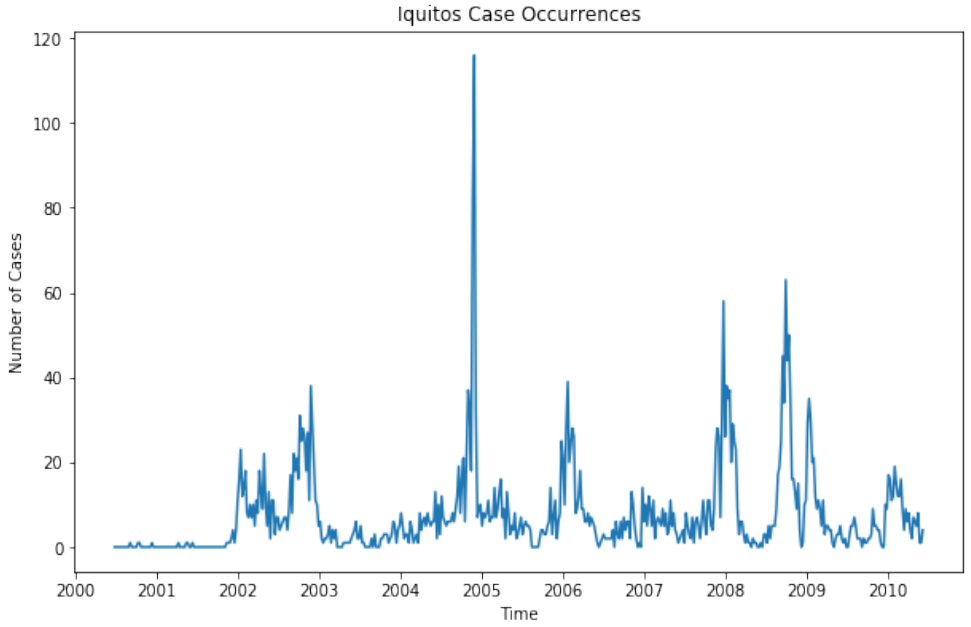
Mean absolute error

$$MAE = \frac{\sum_{n=1}^{N} \left| x_{predict} - x_{observed} \right|}{N}$$

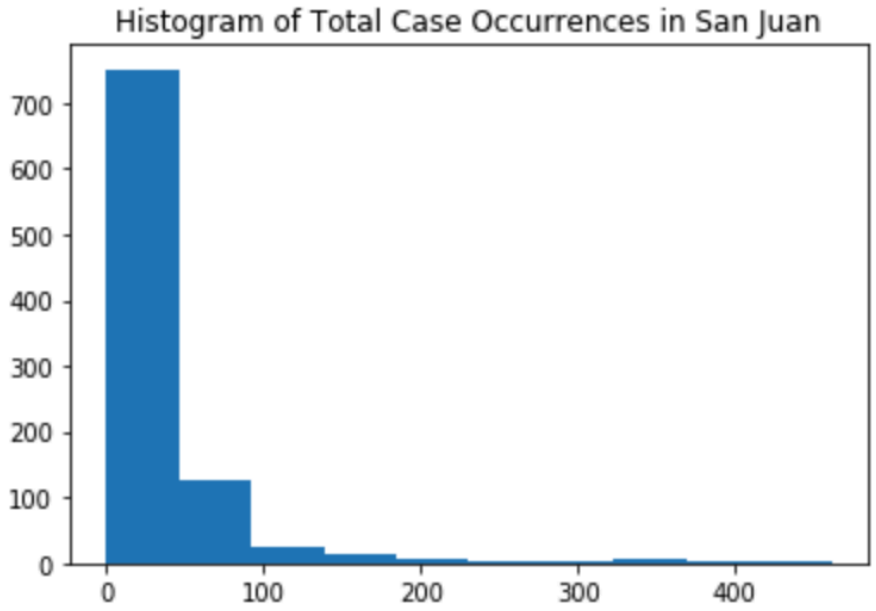
Exploratory Data Analysis

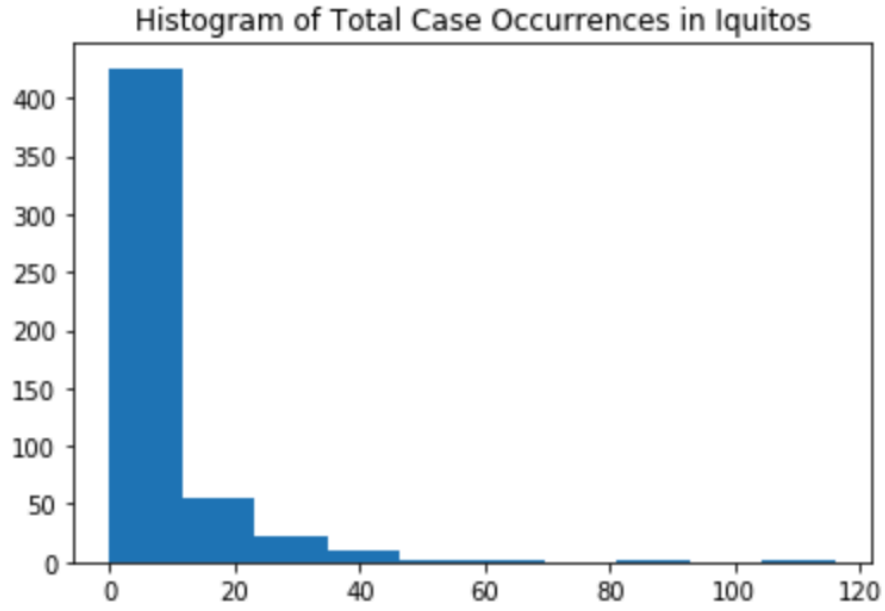
Time-Series Plot: Numbers of Cases



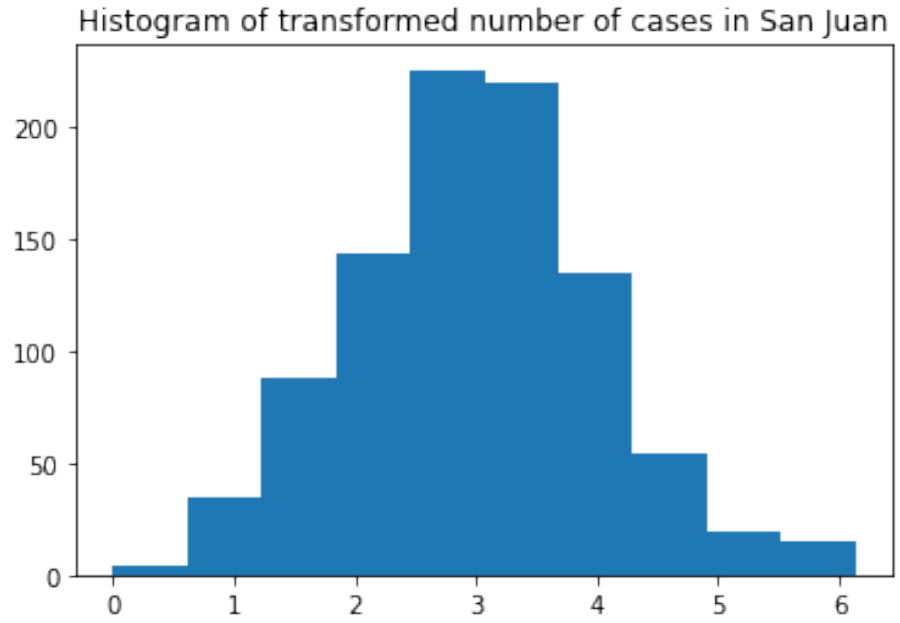


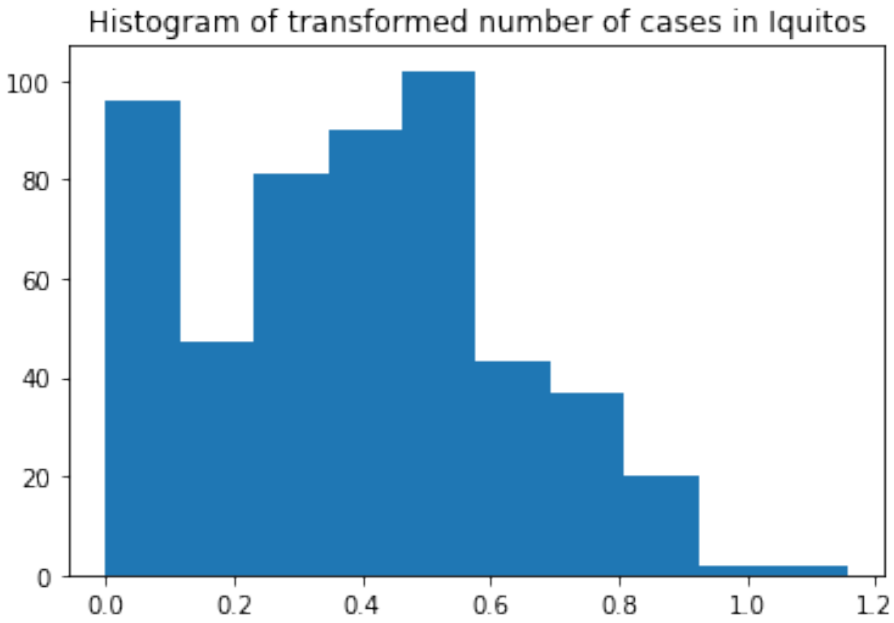
Histograms: Numbers of Cases





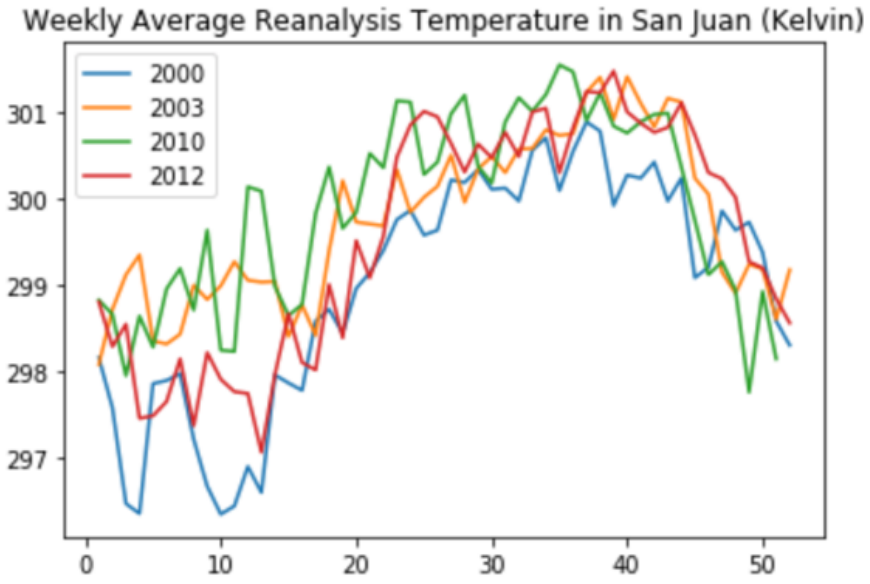
Histograms: Transformed Numbers of Cases

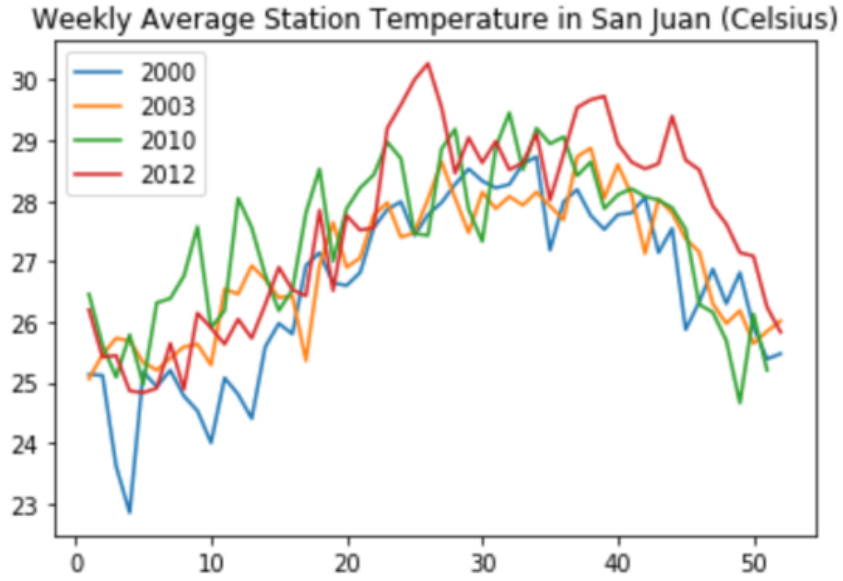




Missing Value

Replace with weekly mean





Modeling

Lag Selection

Week	X	Y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
…	…	…
n	x_n	y_n

Lag 1

X	Y
x_1	y_2
x_2	y_3
x_3	y_4
…	…
X_{n-1}	y_n

Lag 2

X	Y
x_1	y_3
x_2	y_4
x_3	y_5
…	…
X_{n-2}	y_n

Feature Selection: Stepwise Elimination

Features
Precipitation
Average Temperature
Temperature Range
NDVI

Features
Precipitation
~~Average Temperature~~
Temperature Range
~~NDVI~~

Models

Two groups of forecasting techniques will be used:

1. The Autoregressive Integrated Moving Average (ARIMA)
2. Ten Machine Learning algorithms separated into four subgroups:
    a. Neural Network
    b. Regularization Models: LASSO, Bayesian Ridge, Kernel Ridge, and Elastic-Net Regression
    c. Ensemble Models: Random Forest, Gradient Boosting, and Extreme Gradient Boosting.
    d. Robust Regression Models: Huber Regression, and RANSAC Regression

Hyperparameters Tuning: GridSearch

Hyperparameter	Value #1	Value #2
Max Depth	3	4
Number of Estimators	100	500

Combination
[3, 100]
[4, 100]
[3, 500]
[4, 500]

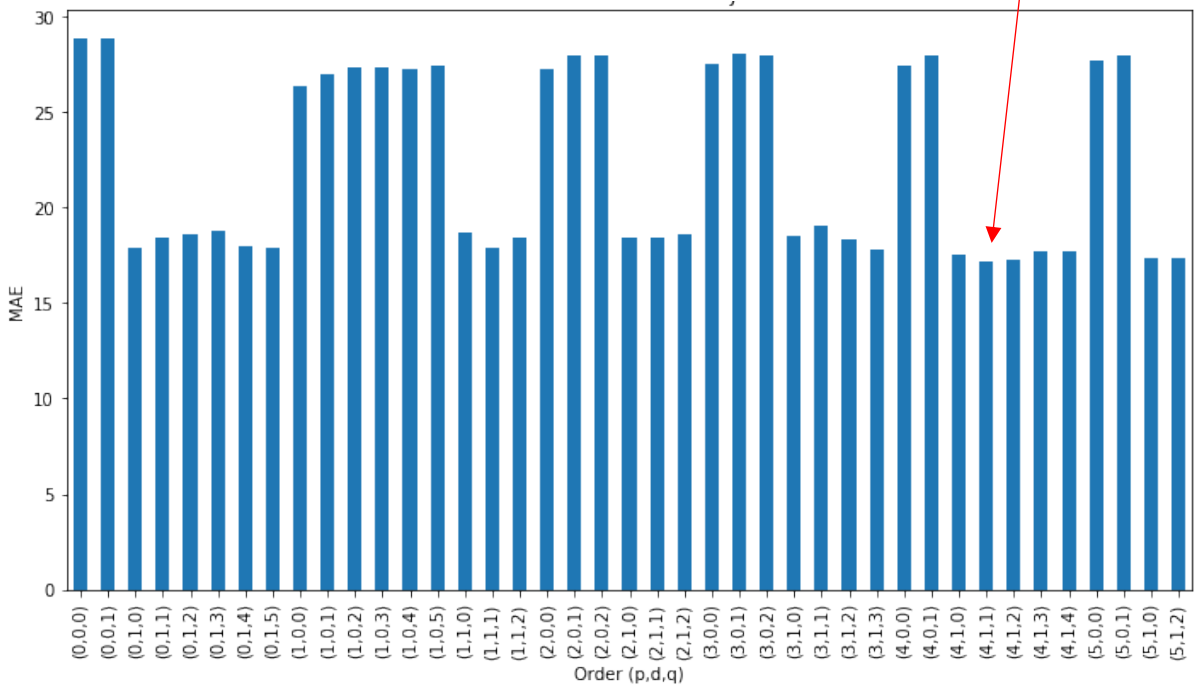
Results

CHULA ENGINEERING
Foundation toward Innovation

Results: ARIMA

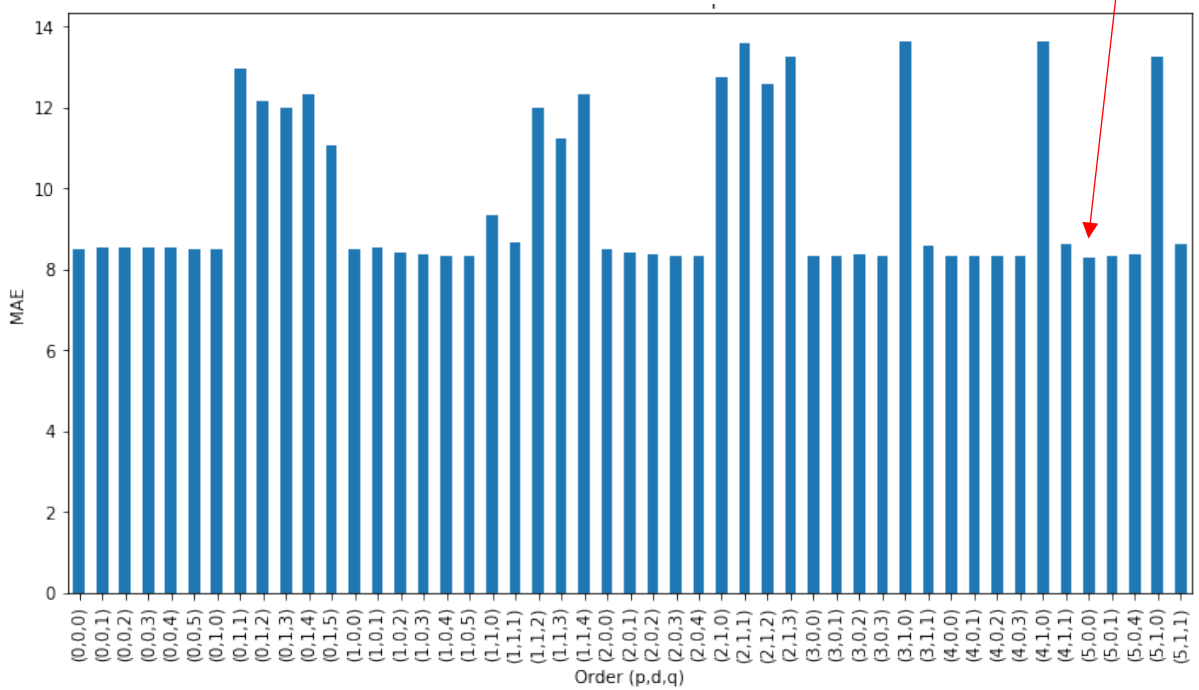
MAE: San Juan

MAE = 17.18



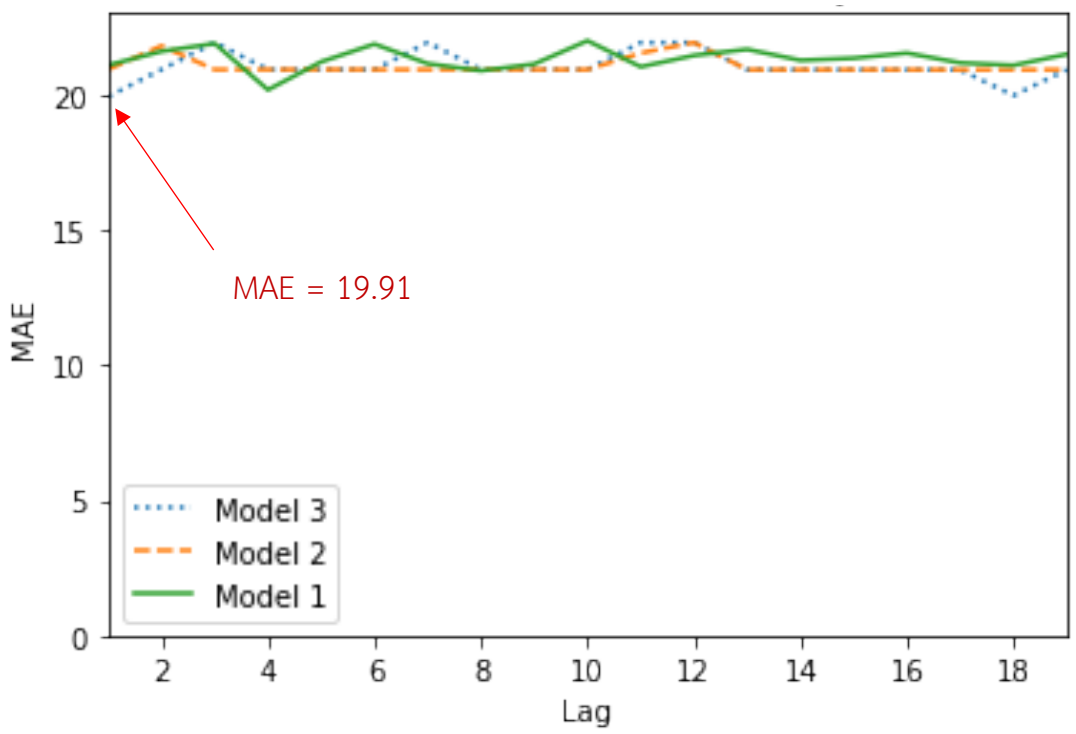
MAE: Iquitos

MAE = 8.29

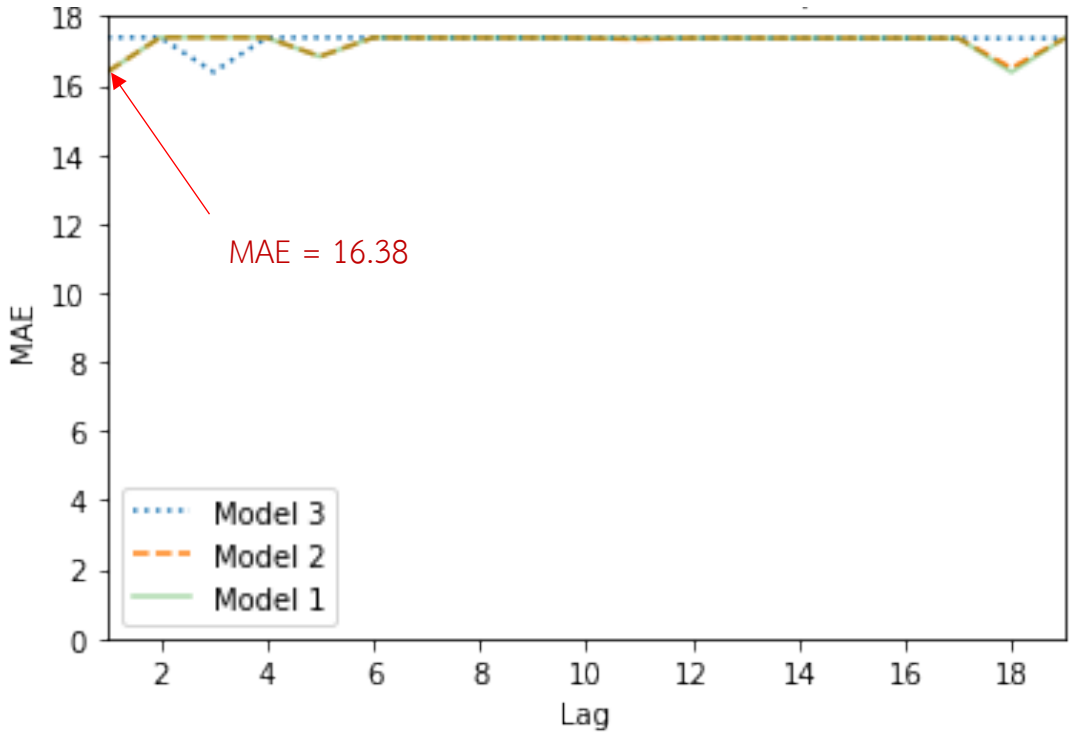


Results: Neural Network

MAE: San Juan



MAE: Iquitos

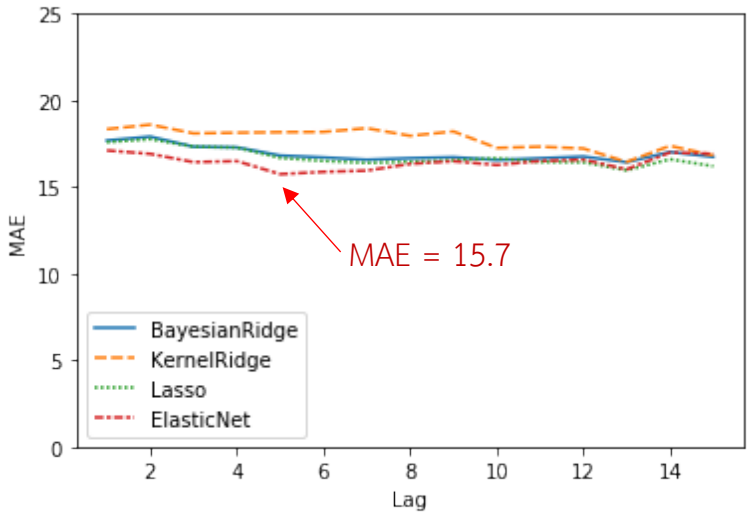


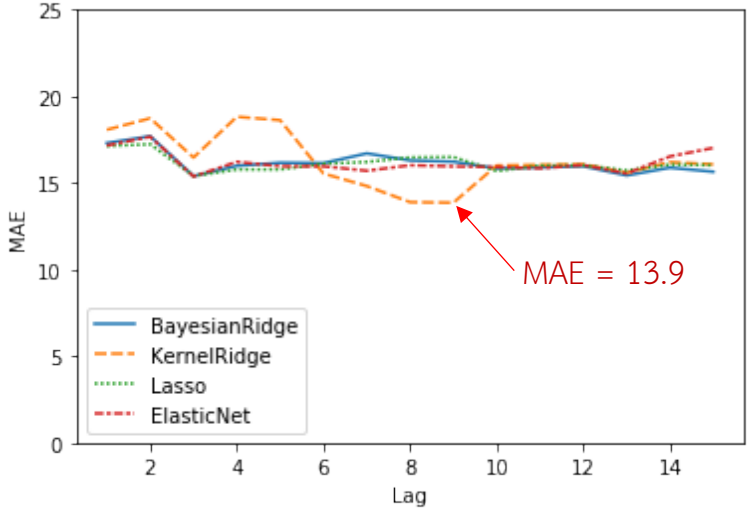
Results: Regularization Models

MAE without Stepwise Elimination

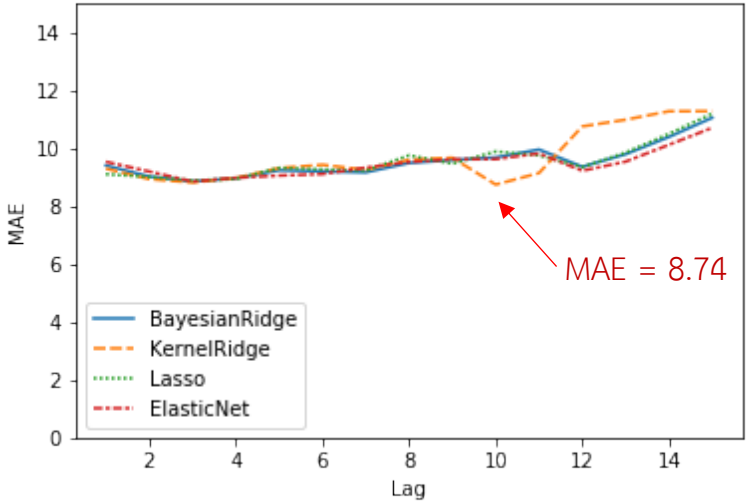
MAE with Stepwise Elimination

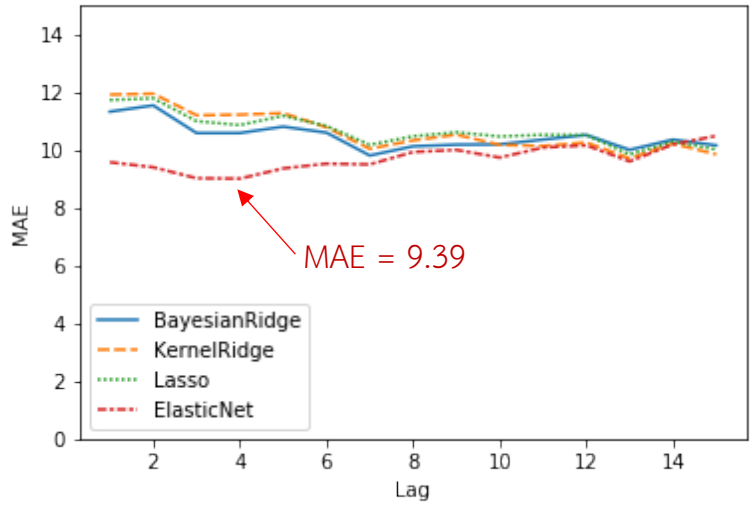
San Juan



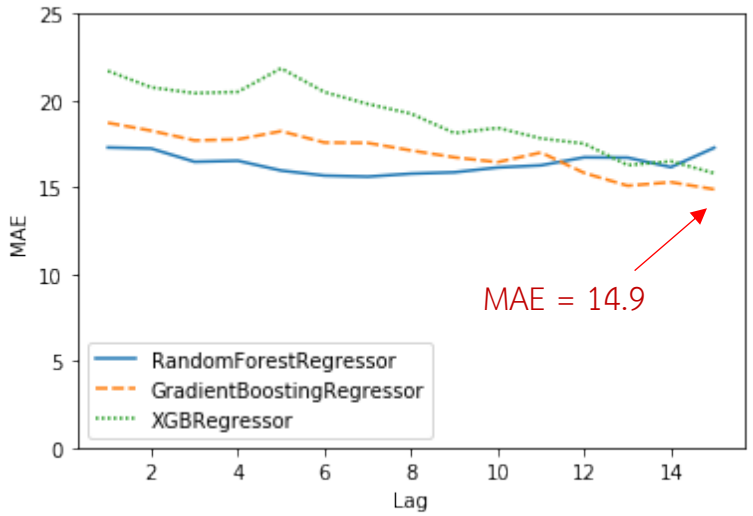
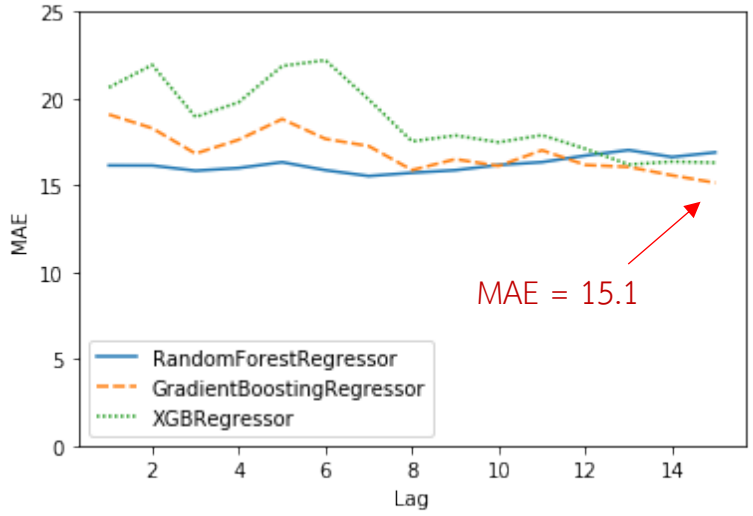
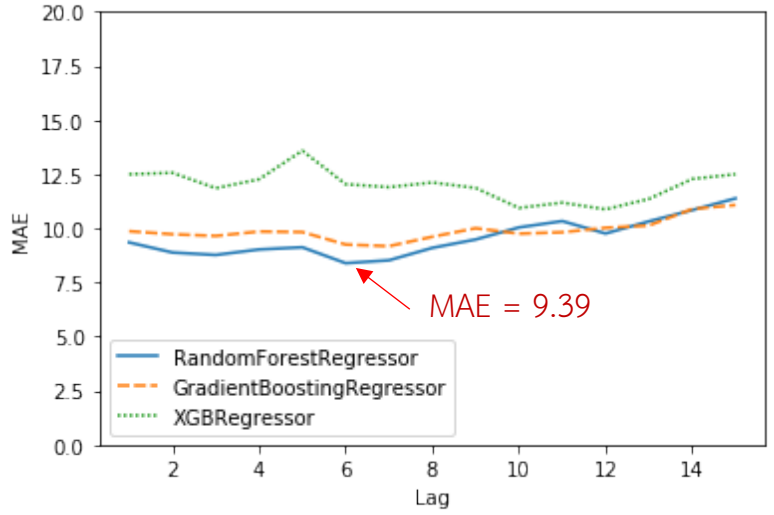
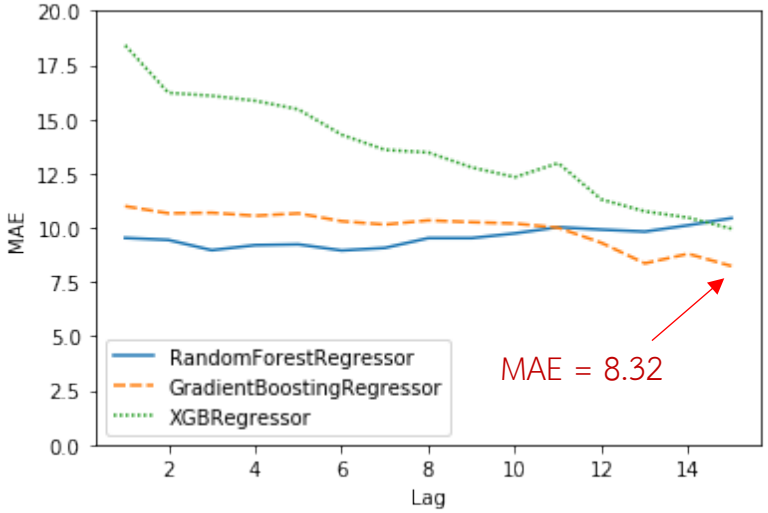


Iquitos





Results: Ensemble Models

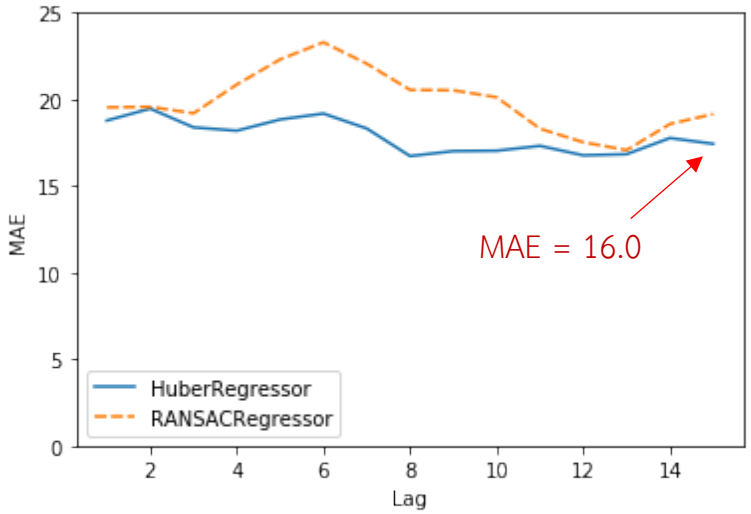
	MAE without Stepwise Elimination	MAE with Stepwise Elimination
San Juan		
Iquitos		

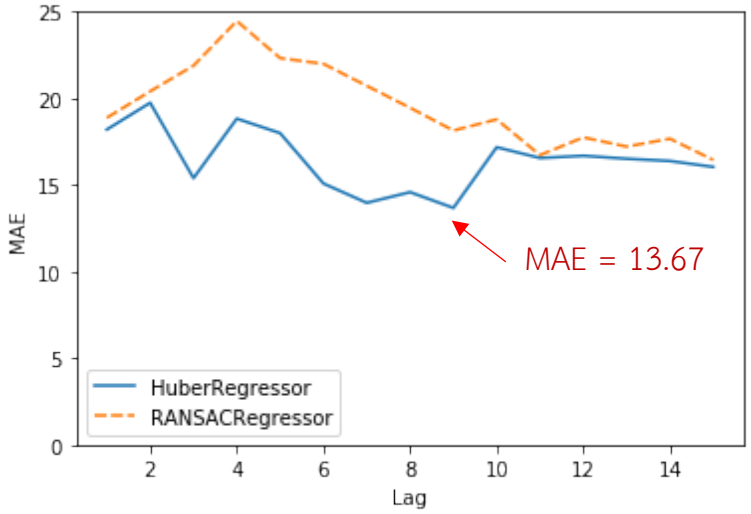
Results: Robust Regression Models

MAE without Stepwise Elimination

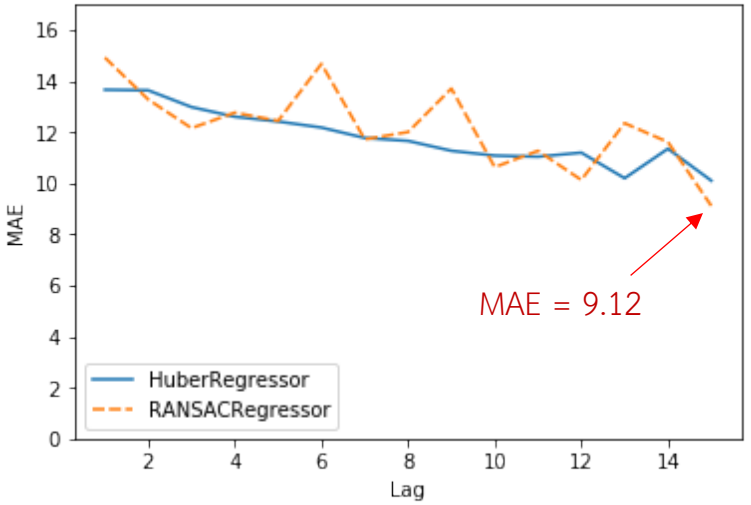
MAE with Stepwise Elimination

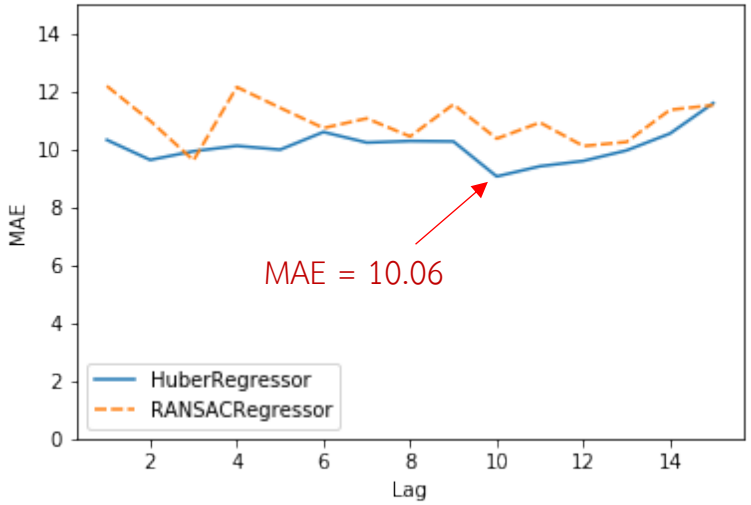
San Juan





Iquitos

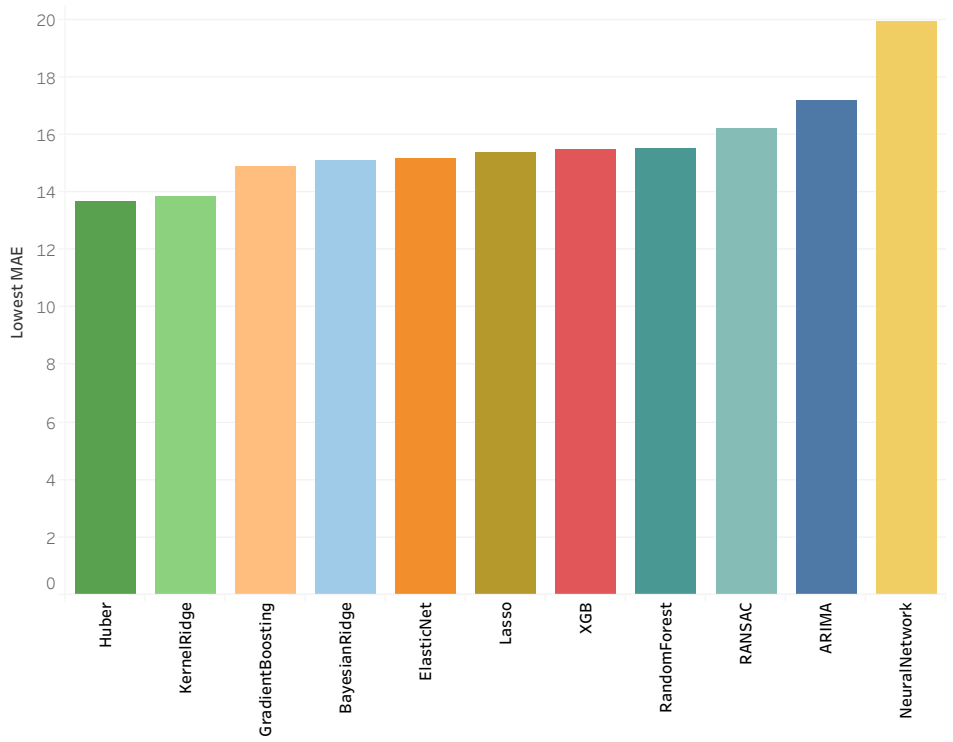




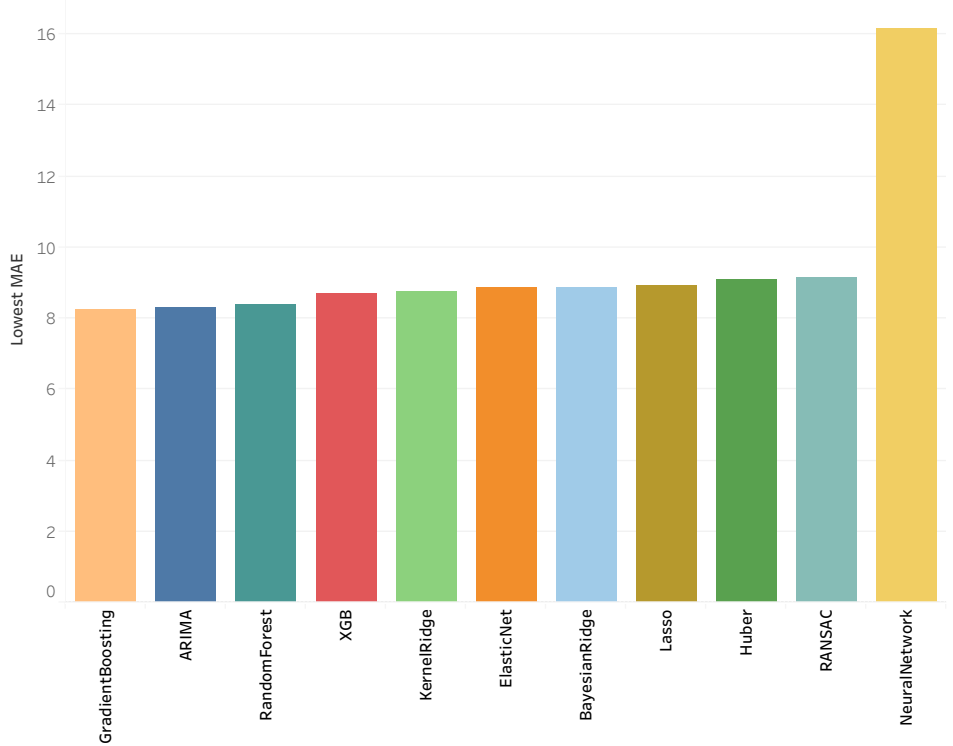
CHULA ENGINEERING
Foundation toward Innovation

Models Summary

San Juan



Iquitos



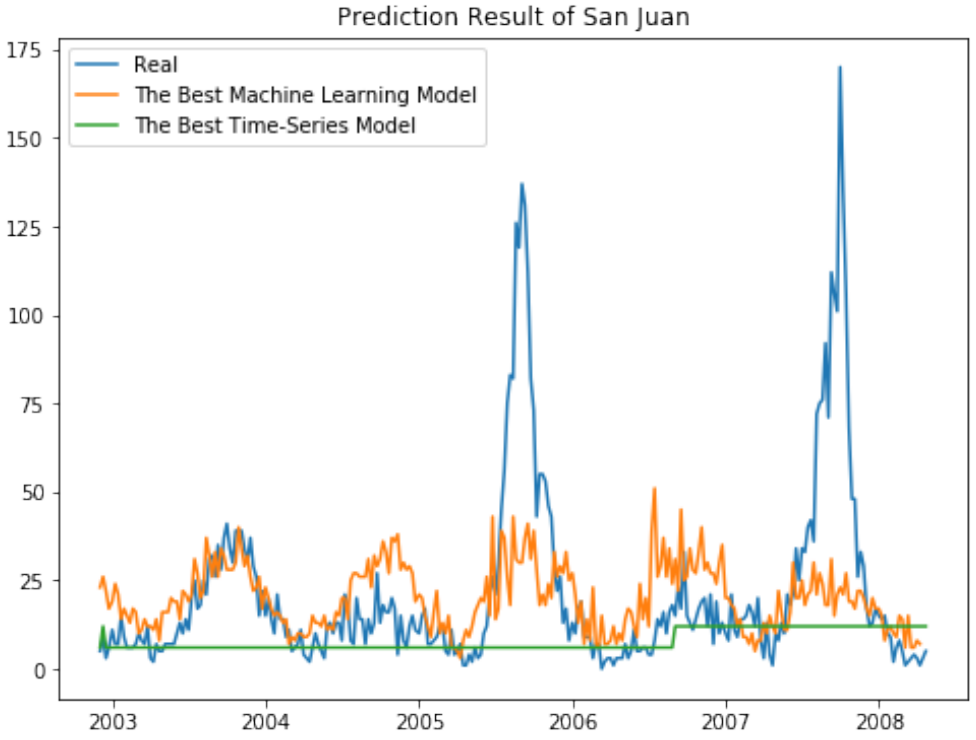
CHULA ENGINEERING
Foundation toward Innovation

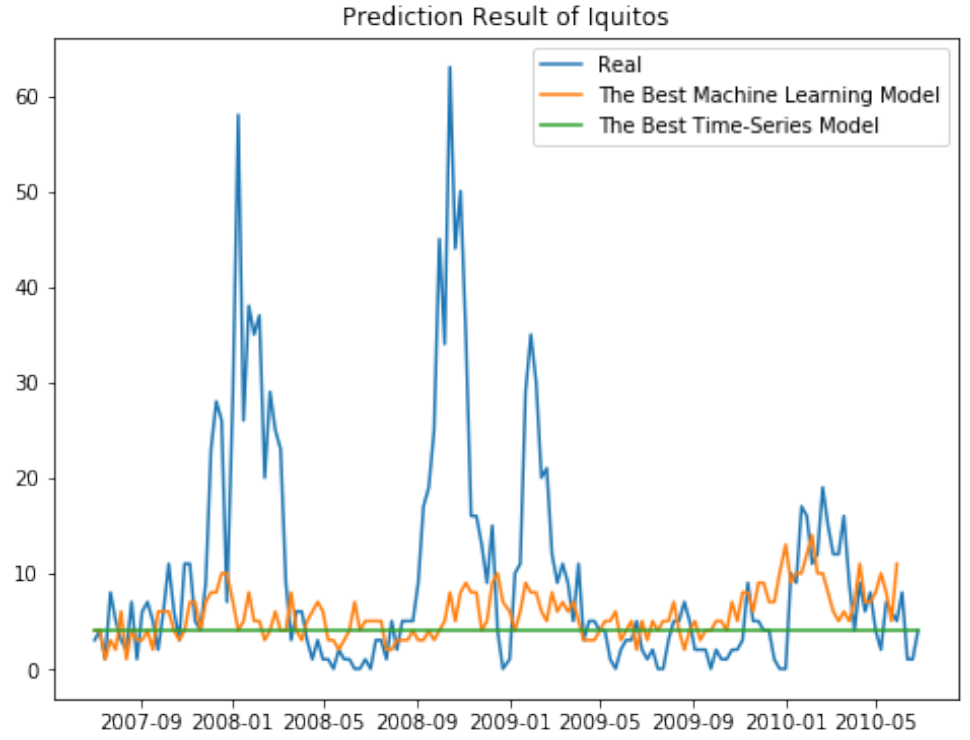
Important Features from Stepwise Elimination

San Juan
NDVI North West
NDVI South East
Reanalysis' Dew Point Temperature
Reanalysis' Maximum Air Temperature
Reanalysis' Relative Humidity
Reanalysis' Diurnal Temperature Range
Station's Average Temperature

Iquitos
NDVI South East
Reanalysis' Average Air Temperature
Reanalysis' Maximum Air Temperature
Reanalysis' Specific Humidity
Reanalysis' Relative Humidity
Reanalysis' Diurnal Temperature Range
Station's Average Temperature

The Best Models' Prediction





Conclusion

Conclusion

- San Juan: Huber Regressor with lag 9 and stepwise elimination, yielding mean absolute error of 13.67
- Iquitos: Gradient Boosting Regressor with lag 15 and stepwise elimination, yielding mean absolute error of 8.23

Thank You