PURDUE UNIVERSITY

Predicting Dengue Fever Cases

Using Machine Learning Methods

Predictive Modeling Project

By: Emily Johncox, Yidan Nie, Yung-Ching Chen
"Modeling Masters"

PURDUE INDUSTRIAL ENGINEERING

PURDUE UNIVERSITY
Environmental and Ecological Engineering

PURDUE UNIVERSITY

Outline

1. **Project Background**
2. **Data Manipulation**
   a. EDA
   b. Cleaning, Imputation, Standardization
3. **Model Development**
   a. Test/Train Data & CV
   b. Tuning parameters for each model
4. **Results**
   a. MAE/RMSE for each model on test
   b. Explanation of best model
5. **Conclusions**

Project Background

What is dengue fever?

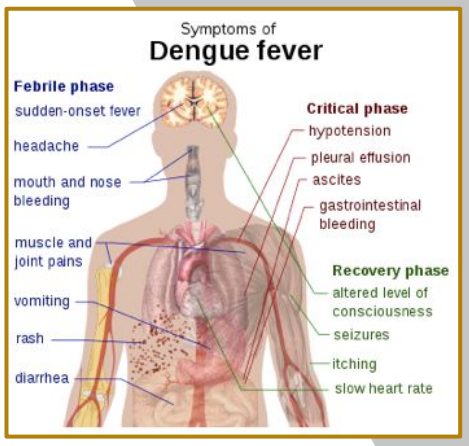
- Mosquito-borne illness
- Symptoms are similar to that of the flu, but in severe cases, can result in death

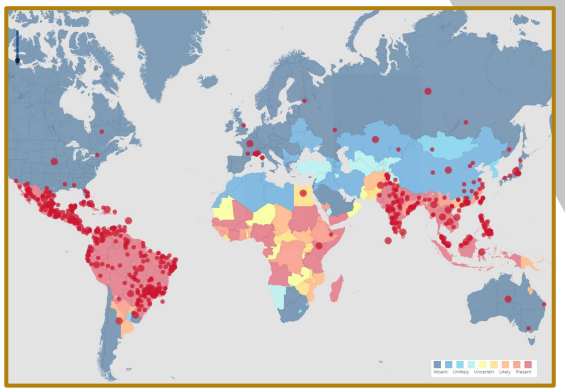
Who is at risk?

- ⅓ of the world's population lives in an at-risk region, mostly the tropics & sub-tropics
- 400 million cases (and 25K deaths) annually

Can dengue fever be prevented or treated?

- No vaccine - best measure is to prevent bites and begin treatment promptly





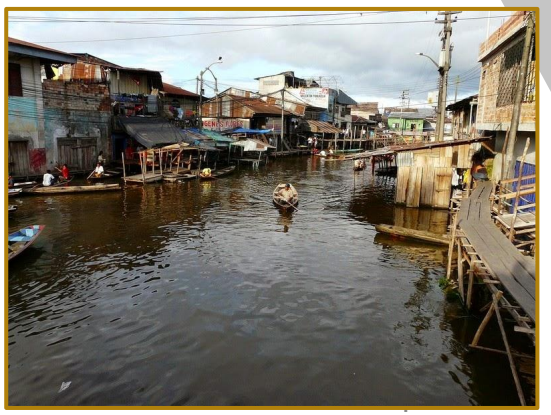
Objective & Hypothesis

Objective

- Develop a model to predict the total number of cases of dengue fever in 2 cities: San Juan, Puerto Rico & Iquitos, Peru
  - Predicting dengue outbreaks allows for public health authorities to better prevent, minimize, and treat dengue fever

Hypothesis

- Because dengue is mosquito-borne, we hypothesize that dengue fever cases will be heavily driven by temperature & precipitation
- Seasonal variation is also expected



Dataset Introduction

Time
- `weekofyear`

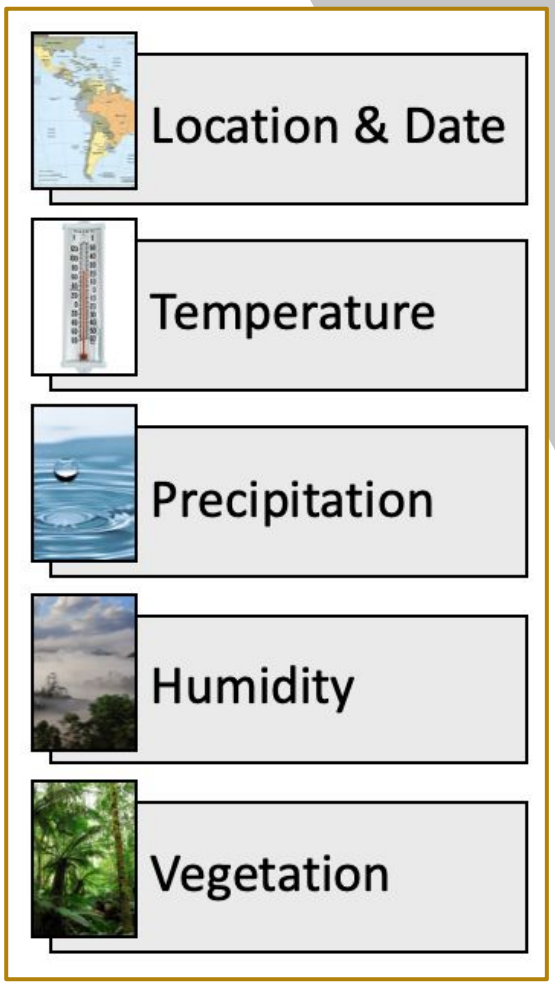
City indicator
- `city` – City abbreviations: `sj` for San Juan and `iq` for Iquitos

Daily climate data weather station measurements
- `station_max_temp_c` – Maximum temperature
- `station_min_temp_c` – Minimum temperature
- `station_avg_temp_c` – Average temperature
- `station_precip_mm` – Total precipitation
- `station_diur_temp_rng_c` – Diurnal temperature range

Satellite precipitation measurements
- `precipitation_amt_mm` – Total precipitation

Climate forecast system reanalysis measurements
- `reanalysis_sat_precip_amt_mm` – Total precipitation
- `reanalysis_dew_point_temp_k` – Mean dew point temperature
- `reanalysis_air_temp_k` – Mean air temperature
- `reanalysis_relative_humidity_percent` – Mean relative humidity
- `reanalysis_specific_humidity_g_per_kg` – Mean specific humidity
- `reanalysis_precip_amt_kg_per_m2` – Total precipitation
- `reanalysis_max_air_temp_k` – Maximum air temperature
- `reanalysis_min_air_temp_k` – Minimum air temperature
- `reanalysis_avg_temp_k` – Average air temperature
- `reanalysis_tdtr_k` – Diurnal temperature range

Satellite vegetation
- `ndvi_se` – Pixel southeast of city centroid
- `ndvi_sw` – Pixel southwest of city centroid
- `ndvi_ne` – Pixel northeast of city centroid
- `ndvi_nw` – Pixel northwest of city centroid

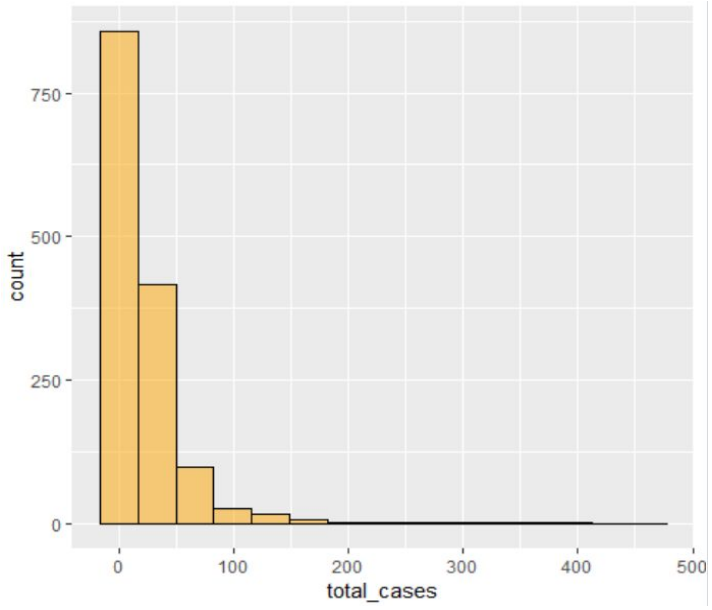


EDA - Response Distribution

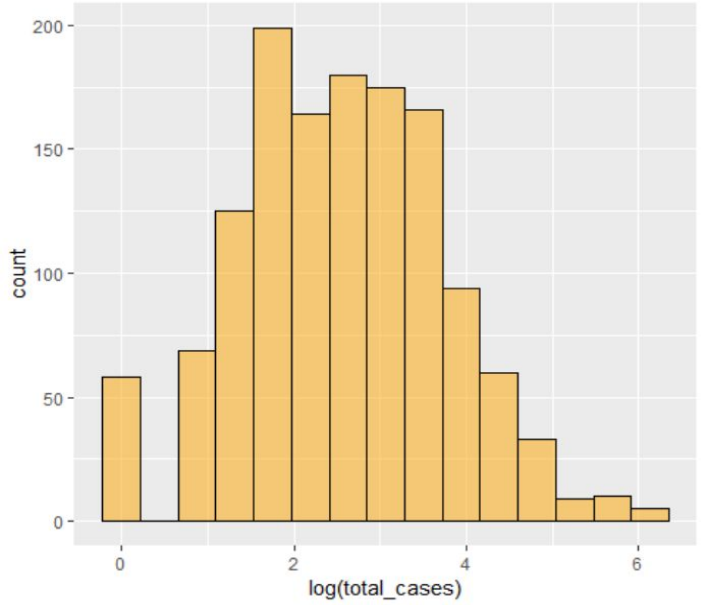
Response variable: **total_cases**

Min	Median	Mean	Max
0	12	24.66	461

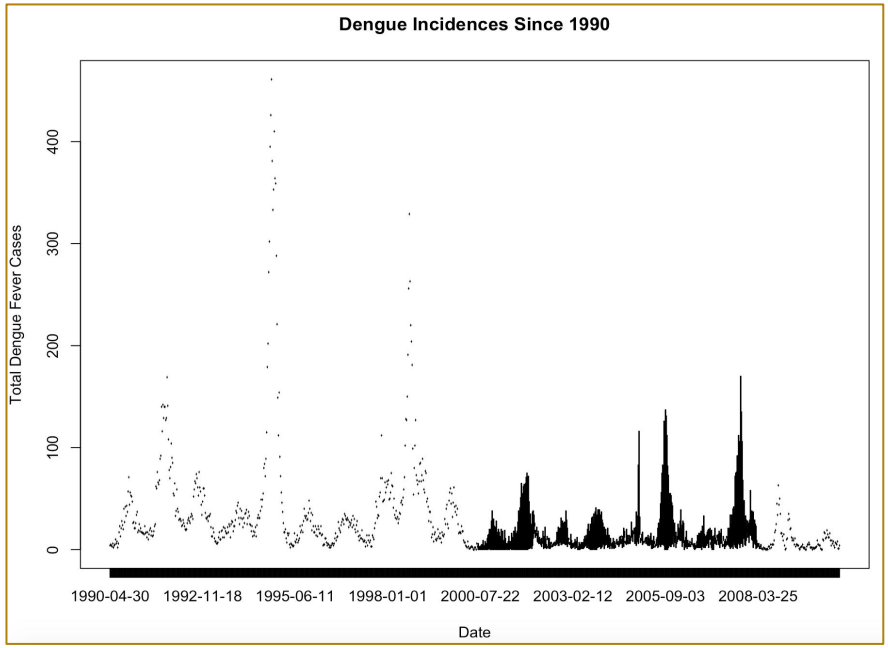
Original

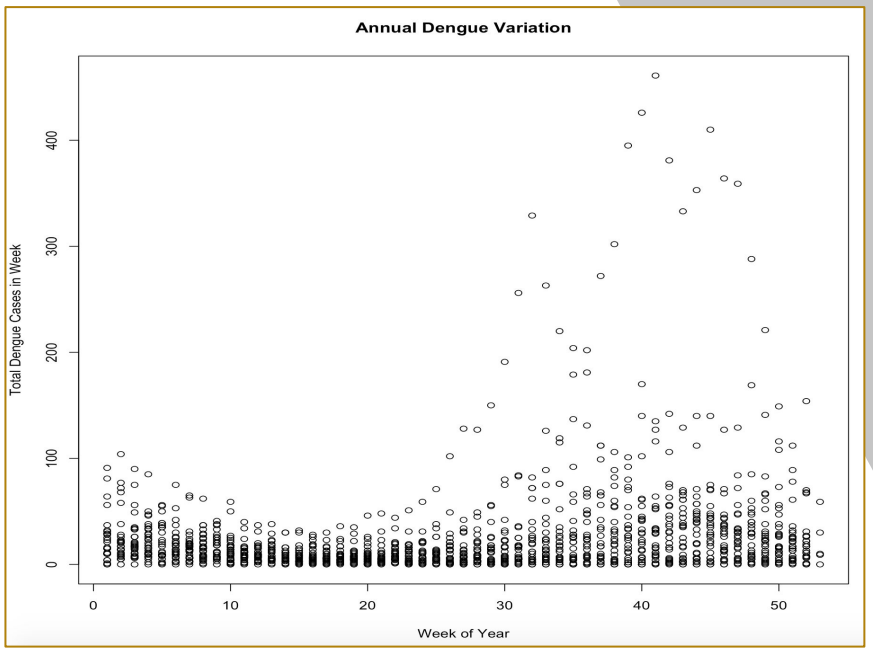


After Transformation



EDA - Dengue vs. Time

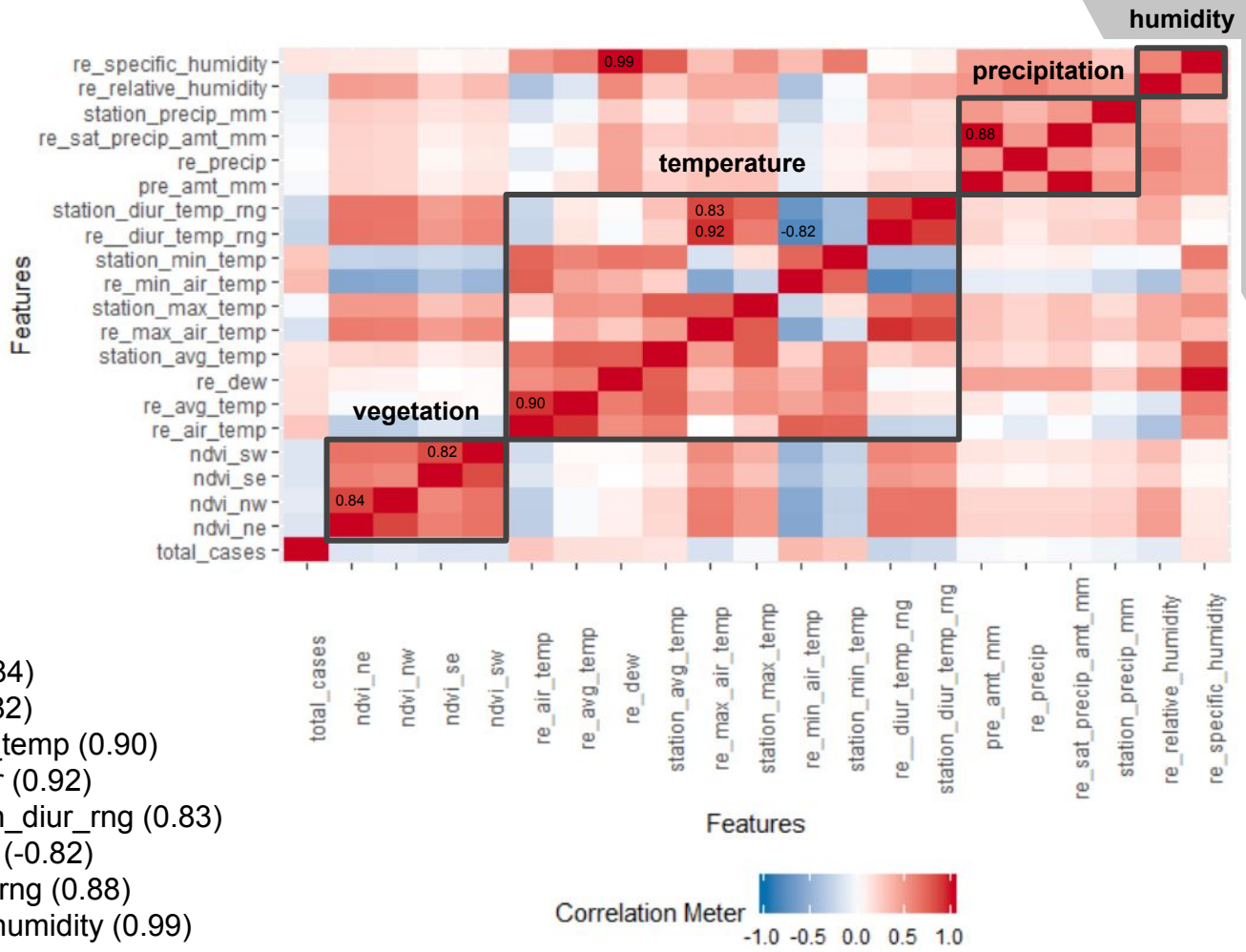




These plots demonstrate a relationship between dengue fever and time.

- *Seasonal variation: Dengue cases are strongly dependent upon the life cycle of mosquitoes, which increase in population in the winter and spring and decrease over the summer as heat increases.*

EDA - Correlation Analysis



High correlation pairs:

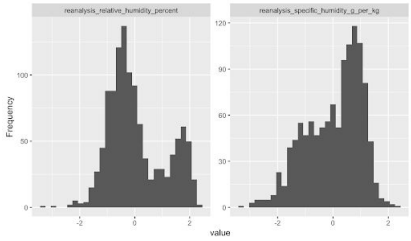
- ndvi_nw, ndvi_ne (0.84)
- ndvi_se, ndvi_sw (0.82)
- re_air_temp, re_avg_temp (0.90)
- re_max_temp, re_tdtr (0.92)
- re_max_temp, station_diur_rng (0.83)
- re_min_temp, re_tdtr (-0.82)
- re_tdte, station_diur_rng (0.88)
- re_dew, re_specific_humidity (0.99)

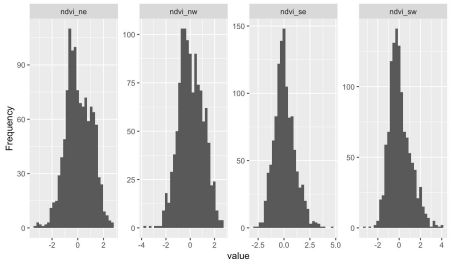
Data Manipulation

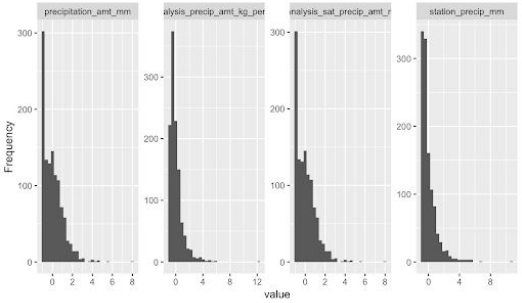
Remove rows with too many NAs (+50%)

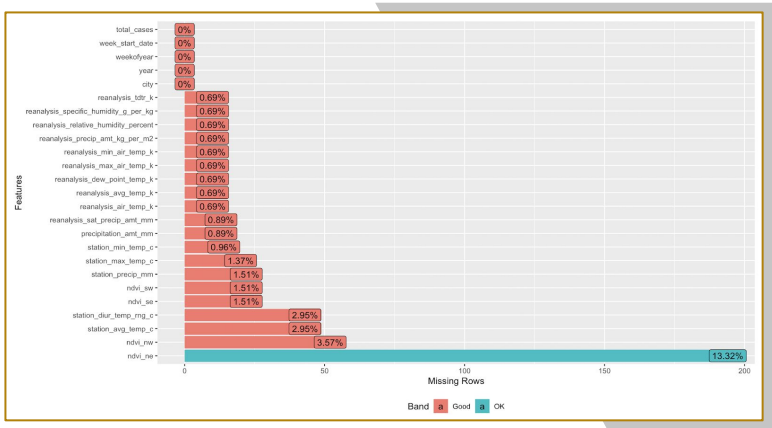
Impute NA values by comparison to prior week

Standardize predictor values

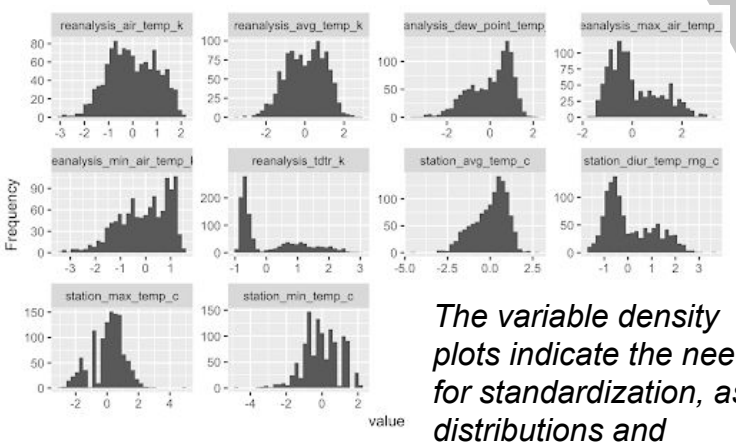








The missing values plot above indicates the need for variable imputation, but that there are no predictors with excessive missing values.



The variable density plots indicate the need for standardization, as distributions and ranges vary greatly.

Model Development Process

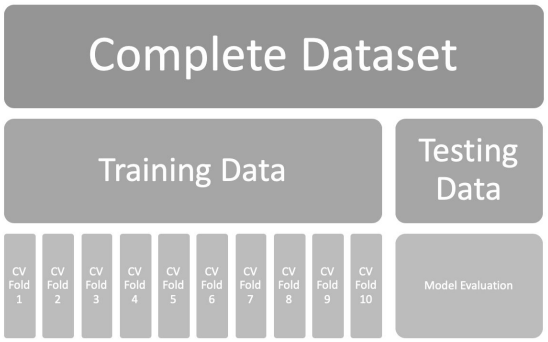
Split dataset:
80% training 20% test set

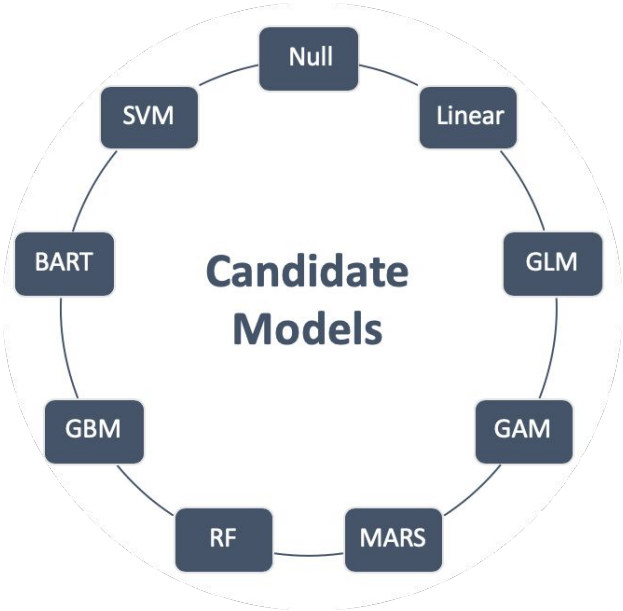
Create a default model of each type

Perform 10-fold CV to tune hyperparameters

Choose optimal hyperparameters using MAE from CV folds

Compare each tuned candidate model using out-of-sample MAE & RMSE





Hyperparameter Tuning

Model	Hyperparameters	Optimal Values
Null	---	---
Linear	Variable selection	See below *
GLM	alpha	0 (ridge)
GAM	Degree of smooth functions	set
MARS	degree	1
	nprune	18
RF	mtry	21
	ntree	750
GBM	depth	7
	shrinkage	0.01
	num.trees	4986
	min_nodes	5
BART	num_tree	50
	k	5
	nu	10
	q	0.75
SVM	cost	0.001

* Optimal Variables in LM:
- Diurnal Temp. Range (Sat)
- Week of Year
- Average Air Temp. (Sat)
- SE Vegetation
- NE Vegetation
- Max. Air Temp. (Sat)
- Diurnal Temp. Range (Sta)

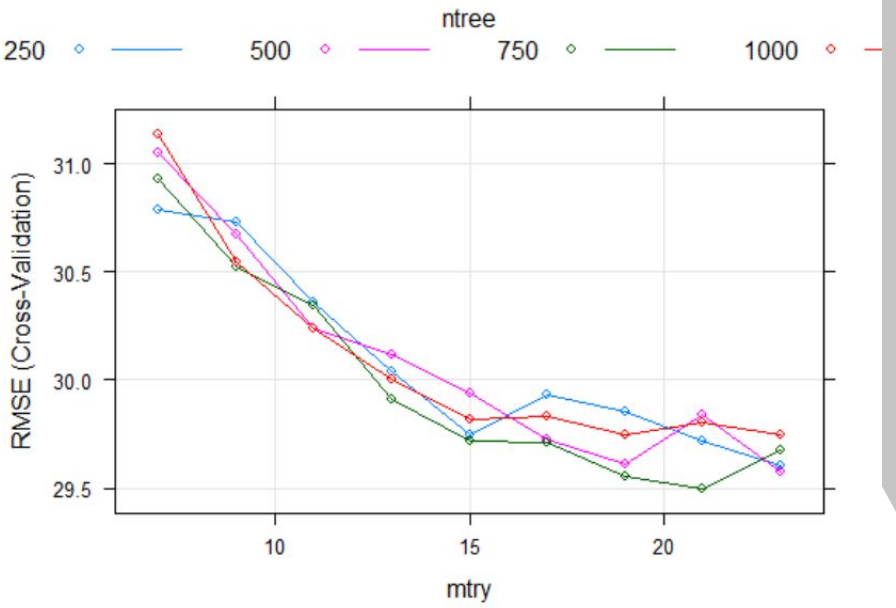
PURDUE UNIVERSITY

Model Results

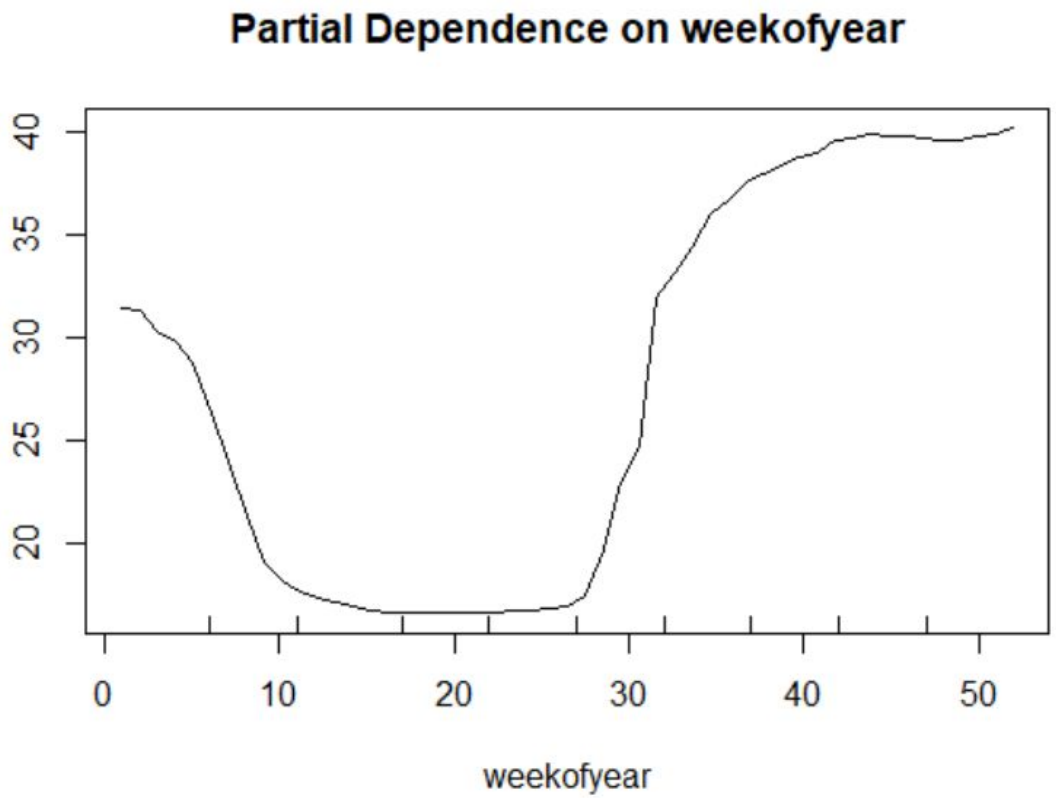
Model	In-Sample Error		Out-of-Sample Error	
	MAE	RMSE	MAE	RMSE
Null	23.87	45.67	21.84	34.65
Linear	17.94	44.03	14.84	32.13
GLM	22.18	41.36	20.05	31.81
GAM	23.26	50.67	18.03	38.16
MARS	17.05	29.64	16.85	27.29
RF	6.98	13.09	14.86	25.96
GBM	7.53	10.42	14.14	26.14
BART	11.88	15.57	18.98	26.61
SVM	19.78	46.63	15.87	34.27

Optimal Model Selection

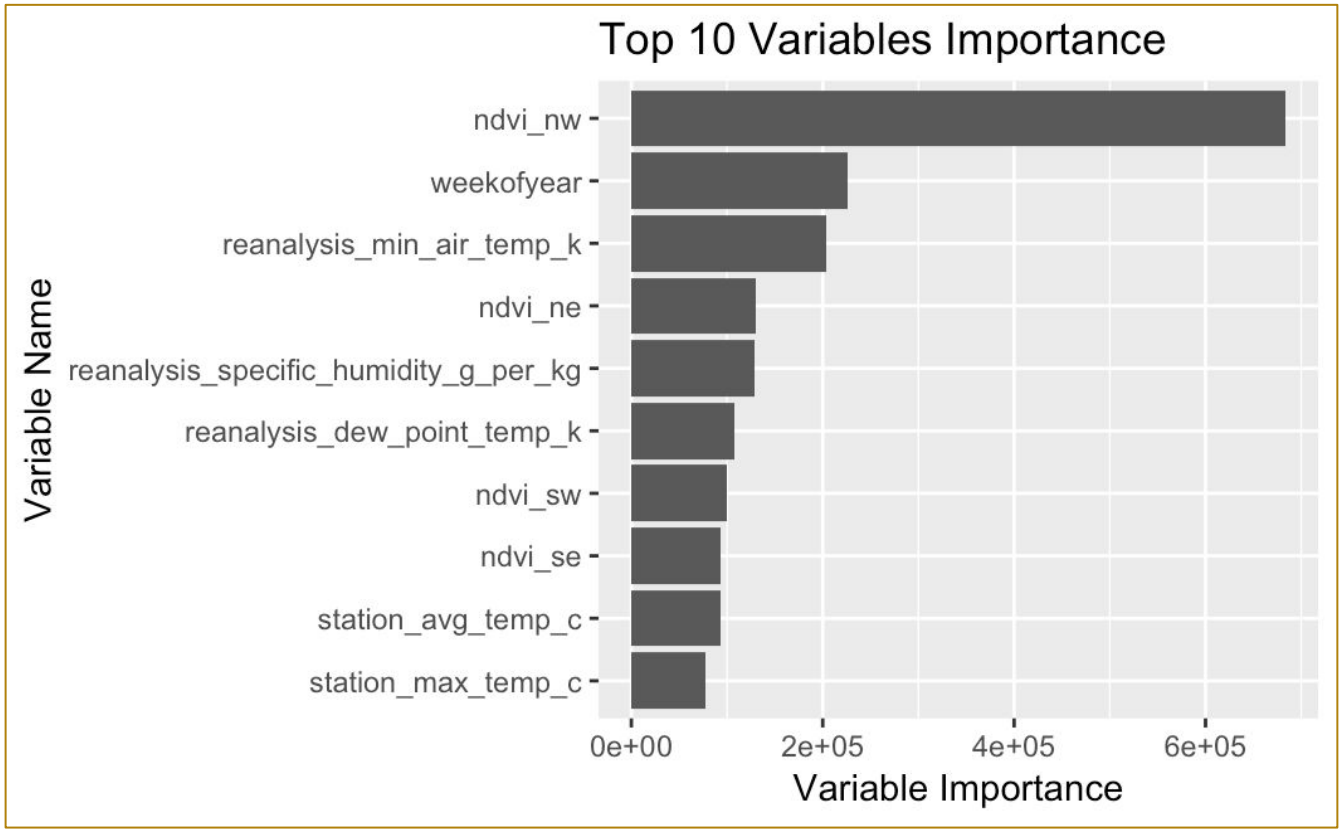
- The random forest (RF) model was selected for its low MAE and RMSE and for its interpretability compared to GBM, which had similar MAE and RMSE values
- The RF model provides an 35% improvement over the MAE of the null model
- Variable importance plots and partial dependence plots can be used to make inferences from this model



This plot demonstrates change in RMSE values as the model was tuned. The optimal condition occurs where mtry = 21 and ntree = 750.

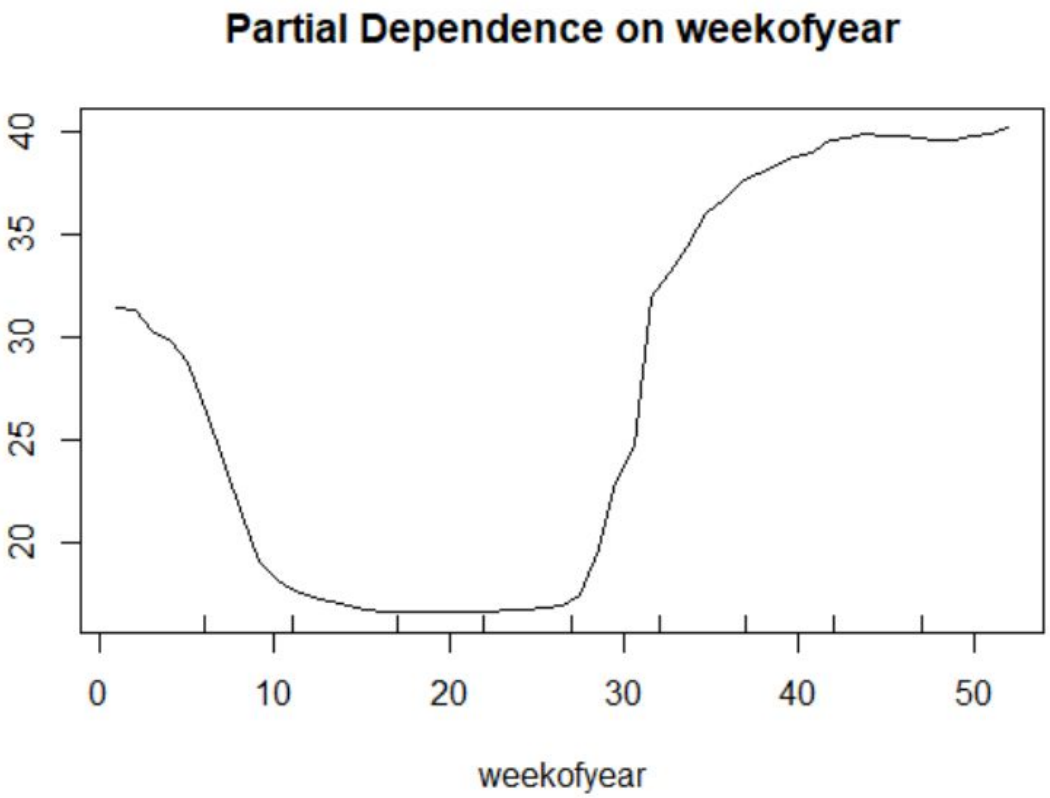

Partial Dependence on weekofyear
20
25
30
35
40
0
10
20
30
40
50
weekofyear

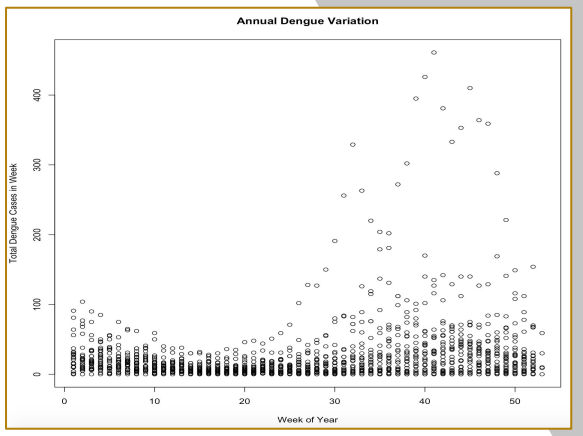
Optimal Model Interpretation



- Vegetation is an important predictor, possibly due to relationships between jungles and mosquito breeding
- Week of year is important to capture seasonal variation
- Temperature is also a key variable

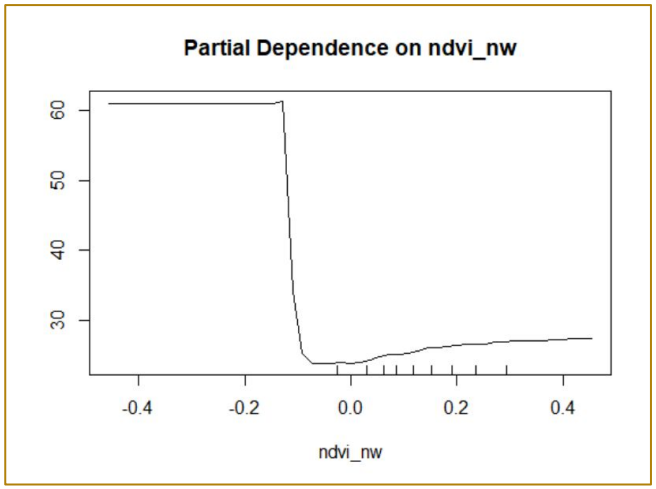
Partial Dependence on Time

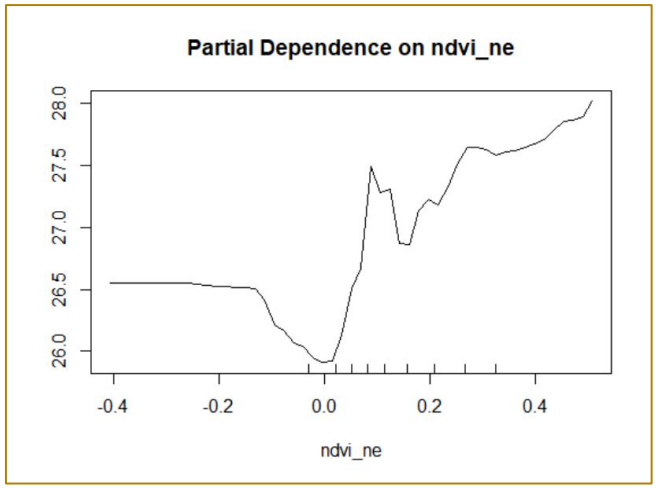


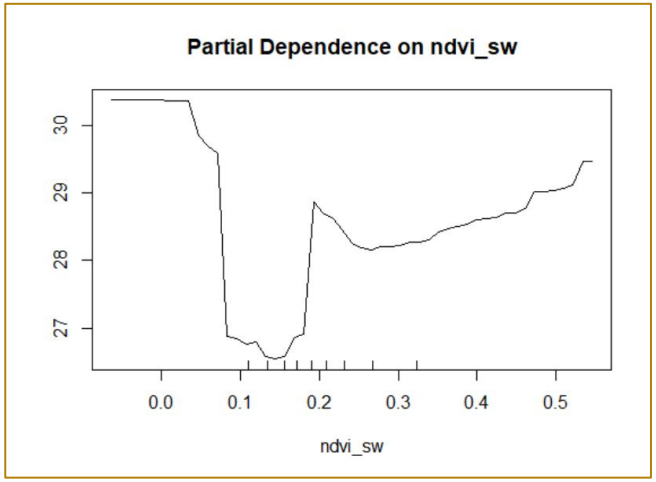


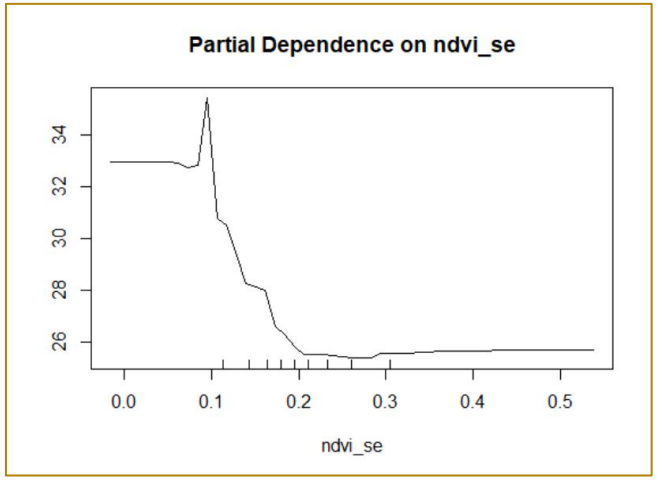
As hypothesized, there is seasonal variation in dengue fever cases with more cases in the spring and summer than in the winter and fall.

Partial Dependence on Vegetation



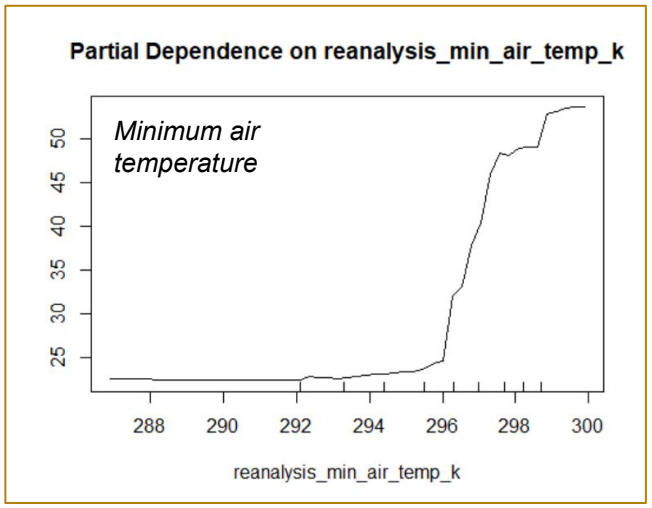


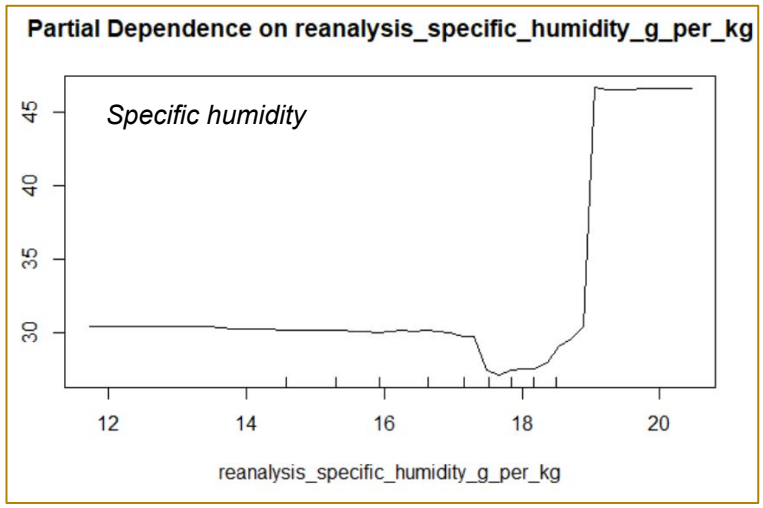


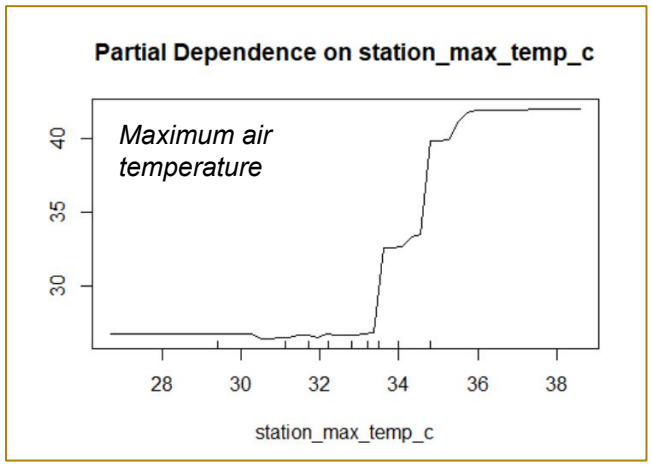


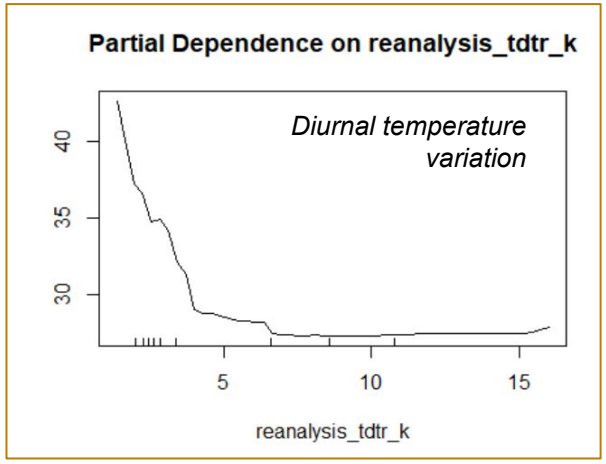
Vegetation dependencies are largely related to location geography and seasonal variation.

Partial Dependence on Temperature & Humidity









There are more dengue cases in warm seasons with low diurnal temperature variation. This corresponds to peak mosquito season.

Conclusions

- Dengue fever can best be predicted using a Random Forest Model
- 35% improvement over null
- Key predictors for dengue fever include **vegetation, season,** and **temperature**
  - This agrees with our hypothesis that there will be annual variation in dengue cases, and that temperature will be an important predictor
  - However, in contrast with our hypothesis, vegetation was identified as a key predictor

Q & A

Thanks!