CS 6375.501 – MACHINE LEARNING

DengAI – Disease Spread Prediction

A

Project Report

submitted

in

partial fulfilment of

Master of Science in Computer Science

April 30 2017

Shiva Podugu (sxp170130)
Manohar Katam (mxk164930)
Sravani Lingam (sxl170330)
Venkata Kartheek Madhavarapu (vxm153830)

I. INTRODUCTION AND DATA EXPLORATION

Our Project is on predicting the **"total_cases"** of Dengue based on **"DengAI: Predicting Disease Spread" dataset** which is an active driven data competitions dataset. Here we have a set of climate information (precipitation, temperature, vegetation) from the two cities: San Juan (sj) and Iquitos (iq) with total cases of dengue count by city, year and week of the year. We aim at making a complete analysis on the DengAI dataset to find the total number of Dengue affected cases in the given two cities with respect to set of climate variables as mentioned above.

The ***DengAI: Predicting Disease Spread*** dataset is taken from an active DrivenData Competition and the link of which is given below:
https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/

BACKGROUND READING:-

Dengue is known to be transmitted seasonally, especially in the rainy season when the creation of stagnant pools allows for more breeding grounds for disease-bearing mosquitoes. Every few years, however, these spikes burst into epidemics, which are still more or less sporadic. Moreover, climate change is expected to permit the entry of disease agents into new territories that are growing increasingly temperate, thus infecting populations as yet unfamiliar with the disease. The unpredictability and scale of dengue causes it to remain an issue of wide concern.

The relationship between dengue transmission and the environment is particularly tenuous, with a complex network of variables and interactions. The study **Climate and Dengue Transmission: Evidence and Implications** (http://ehp.niehs.nih.gov/wp-content/uploads/121/11-12/ehp.1306556.pdf) explains insights into it.

II. DATASET DESCRIPTION

City and date indicators.

1. city: 'sj' for San Juan and 'iq' for Iquitos

2. year:

3. weekofyear:

4. week_start_date: the start date of each week, as given in dd-mm-yyyy format

NOAA's GHCN daily climate data weather station measurements: NOAA is the U.S.' National Oceanic and Atmospheric Association, and the GHCN (or the Global Historical Climatology Network) is their database integrating climate reports across land and sea stations around the world. All temperature values here are in degrees Celsius.

5. station_max_temp_c: Maximum temperature

6. station_min_temp_c: Minimum temperature

7. station_avg_temp_c: Average temperature,

8. station_precip_mm: Total precipitation

9. station_diur_temp_rng_c: Diurnal temperature range

PERSIANN satellite precipitation measurements (0.25x0.25 degree scale): PERSIANN, on the other hand, is the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks, as developed by UC Irvine's Centre for Hydrometeorology and Remote Sensing (CHRS). As its name might suggest, the system uses neural networks to estimate rainfall rate at a given geographic location, so it may be interesting to see how the values here differ from those given by NOAA's different measurements.

10. precipitation_amt_mm: Total precipitation

NOAA's NCEP Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale): NCEP are NOAA's National Centers for Environmental Prediction. At its simplest, the CFS, or the Climate Forecast System, is a model of the interaction among the earth's lands, oceans, and temperature based on hourly data. All temperature values here are in Kelvin.

11. reanalysis_sat_precip_amt_mm: Total precipitation (expressed in millimeters)

12. reanalysis_dew_point_temp_k: Mean dew point temperature (The temperature at which air would have to cool in order to reach saturation)

13. reanalysis_air_temp_k: Mean air temperature

14. reanalysis_relative_humidity_percent: Mean relative humidity (The amount of water vapor in the air, expressed as the percentage of the amount needed for the air to be saturated at the same temperature)

15. reanalysis_specific_humidity_g_per_kg: Mean specific humidity (The amount of water vapor in the air, with respect to the total mass of air + water vapor)

16. reanalysis_precip_amt_kg_per_m2: Total precipitation (expressed as kg per meters squared)

17. reanalysis_max_air_temp_k: Maximum air temperature

18. reanalysis_min_air_temp_k: Minimum air temperature

19. reanalysis_avg_temp_k: Average air temperature

20. reanalysis_tdtr_k: Diurnal temperature range

Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's CDR Normalized Difference Vegetation Index (0.5x0.5 degree scale) measurements: The NDVI is an indicator that measures the presence of green vegetation on a given pixel of land surfaces. This is done by searching for the distinct wavelengths of sunlight absorbed (visible) and reflected (near-infrared) by plants for photosynthesis. The values here range between 0 and 0.8, with NDVI's between 0.3 and 0.8 indicating the presence of vegetation, and those below 0.3 bare soils. I'm expecting that higher vegetation would at least be correlated with higher numbers of cases.

21. ndvi_se – Pixel southeast of city centroid

22. ndvi_sw – Pixel southwest of city centroid

23. ndvi_ne – Pixel northeast of city centroid

24. ndvi_nw – Pixel northwest of city centroid

And lastly the target variable.

25. total_cases: the number of cases within the timeframe for a given city

Total number of attributes = 24 (predictors)
Total number of instances = 1456

Indicators mentioned in “background reading” that aren't covered in the above are *rate of evaporation, ENSO indices, and sea surface temperatures*.

III. PREPROCESSING

Splitting the data into two datasets based on city:

San Juan City:

Iquitos City:

San Juan is in coastal area where as Iquitos is located much inside to the coastal region. So we assume that the occurrence of dengue is different in both regions and we split the data into two datasets based on city attribute.

Missing Values: Read in the data and *summary* command in R gives insight into minimum and maximum values of the attribute and missing values (or NA's) in the attribute.

```
> summary(train)
```

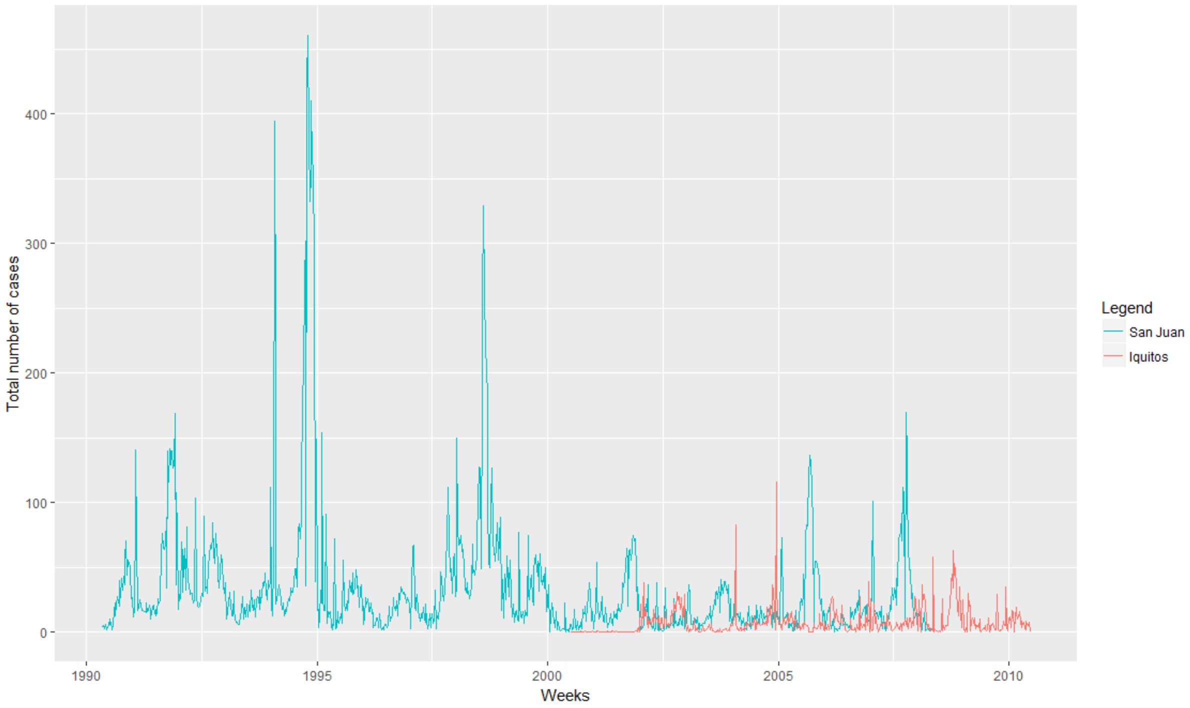
```
city          year        weekofyear     week_start_date     ndvi_ne
 iq:520   Min.   :1990   Min.   : 1.00   01-01-2001:   2   Min.   :-0.4062
 sj:936   1st Qu.:1997   1st Qu.:13.75   01-01-2002:   2   1st Qu.: 0.0449
          Median :2002   Median :26.50   01-01-2003:   2   Median : 0.1288
          Mean   :2001   Mean   :26.50   01-01-2004:   2   Mean   : 0.1421
          3rd Qu.:2005   3rd Qu.:39.25   01-01-2005:   2   3rd Qu.: 0.2485
          Max.   :2010   Max.   :53.00   01-01-2006:   2   Max.   : 0.5084
                                         (Other)   :1444   NA's   :193

ndvi_nw              ndvi_se              ndvi_sw          precipitation_amt_mm
 Min.   :-0.45610   Min.   :-0.01553   Min.   :-0.06346   Min.   :  0.00
 1st Qu.: 0.04922   1st Qu.: 0.15509   1st Qu.: 0.14421   1st Qu.:  9.80
 Median : 0.12143   Median : 0.19605   Median : 0.18945   Median : 38.34
 Mean   : 0.13055   Mean   : 0.20378   Mean   : 0.20231   Mean   : 45.76
 3rd Qu.: 0.21660   3rd Qu.: 0.24885   3rd Qu.: 0.24698   3rd Qu.: 70.23
 Max.   : 0.45443   Max.   : 0.53831   Max.   : 0.54602   Max.   :390.60
 NA's   :52         NA's   :22         NA's   :22         NA's   :13

reanalysis_air_temp_k reanalysis_avg_temp_k reanalysis_dew_point_temp_k
 Min.   :294.6         Min.   :294.9         Min.   :289.6
 1st Qu.:297.7         1st Qu.:298.3         1st Qu.:294.1
 Median :298.6         Median :299.3         Median :295.6
 Mean   :298.7         Mean   :299.2         Mean   :295.2
 3rd Qu.:299.8         3rd Qu.:300.2         3rd Qu.:296.5
```

```
 Max.   :302.2         Max.   :302.9         Max.   :298.4  
 NA's   :10            NA's   :10            NA's   :10     

reanalysis_max_air_temp_k reanalysis_min_air_temp_k reanalysis_precip_amt_
kg_per_m2
 Min.   :297.8             Min.   :286.9             Min.   :  0.00       
 1st Qu.:301.0             1st Qu.:293.9             1st Qu.: 13.05       
 Median :302.4             Median :296.2             Median : 27.25       
 Mean   :303.4             Mean   :295.7             Mean   : 40.15       
 3rd Qu.:305.5             3rd Qu.:297.9             3rd Qu.: 52.20       
 Max.   :314.0             Max.   :299.9             Max.   :570.50       
 NA's   :10                NA's   :10                NA's   :10           

reanalysis_relative_humidity_percent reanalysis_sat_precip_amt_mm
 Min.   :57.79                        Min.   :  0.00              
 1st Qu.:77.18                        1st Qu.:  9.80              
 Median :80.30                        Median : 38.34              
 Mean   :82.16                        Mean   : 45.76              
 3rd Qu.:86.36                        3rd Qu.: 70.23              
 Max.   :98.61                        Max.   :390.60              
 NA's   :10                           NA's   :13                  

reanalysis_specific_humidity_g_per_kg reanalysis_tdtr_k station_avg_temp_c
 Min.   :11.72                         Min.   : 1.357    Min.   :21.40     
 1st Qu.:15.56                         1st Qu.: 2.329    1st Qu.:26.30     
 Median :17.09                         Median : 2.857    Median :27.41     
 Mean   :16.75                         Mean   : 4.904    Mean   :27.19     
 3rd Qu.:17.98                         3rd Qu.: 7.625    3rd Qu.:28.16     
 Max.   :20.46                         Max.   :16.029    Max.   :30.80     
 NA's   :10                            NA's   :10        NA's   :43        
station_diur_temp_rng_c station_max_temp_c station_min_temp_c station_prec
ip_mm
 Min.   : 4.529          Min.   :26.70      Min.   :14.7       Min.   :  0
.00  
 1st Qu.: 6.514          1st Qu.:31.10      1st Qu.:21.1       1st Qu.:  8
.70  
 Median : 7.300          Median :32.80      Median :22.2       Median : 23
.85  
 Mean   : 8.059          Mean   :32.45      Mean   :22.1       Mean   : 39
.33  
 3rd Qu.: 9.567          3rd Qu.:33.90      3rd Qu.:23.3       3rd Qu.: 53
.90  
 Max.   :15.800          Max.   :42.20      Max.   :25.6       Max.   :543
.30  
 NA's   :43              NA's   :20         NA's   :14         NA's   :22     

  total_cases    
 Min.   :  0.00  
 1st Qu.:  5.00  
 Median : 12.00  
 Mean   : 24.68  
 3rd Qu.: 28.00  
 Max.   :461.00  
```

Observation from summary: The count of NA's at the bottom of each variable indicates that there is an underlying pattern to the missing values for each data source. Nevertheless, **since these are variables that are known to follow seasonal trends, we can impute them by taking the most recent values (except for NDVI, because it has clusters of values that are all missing)**.

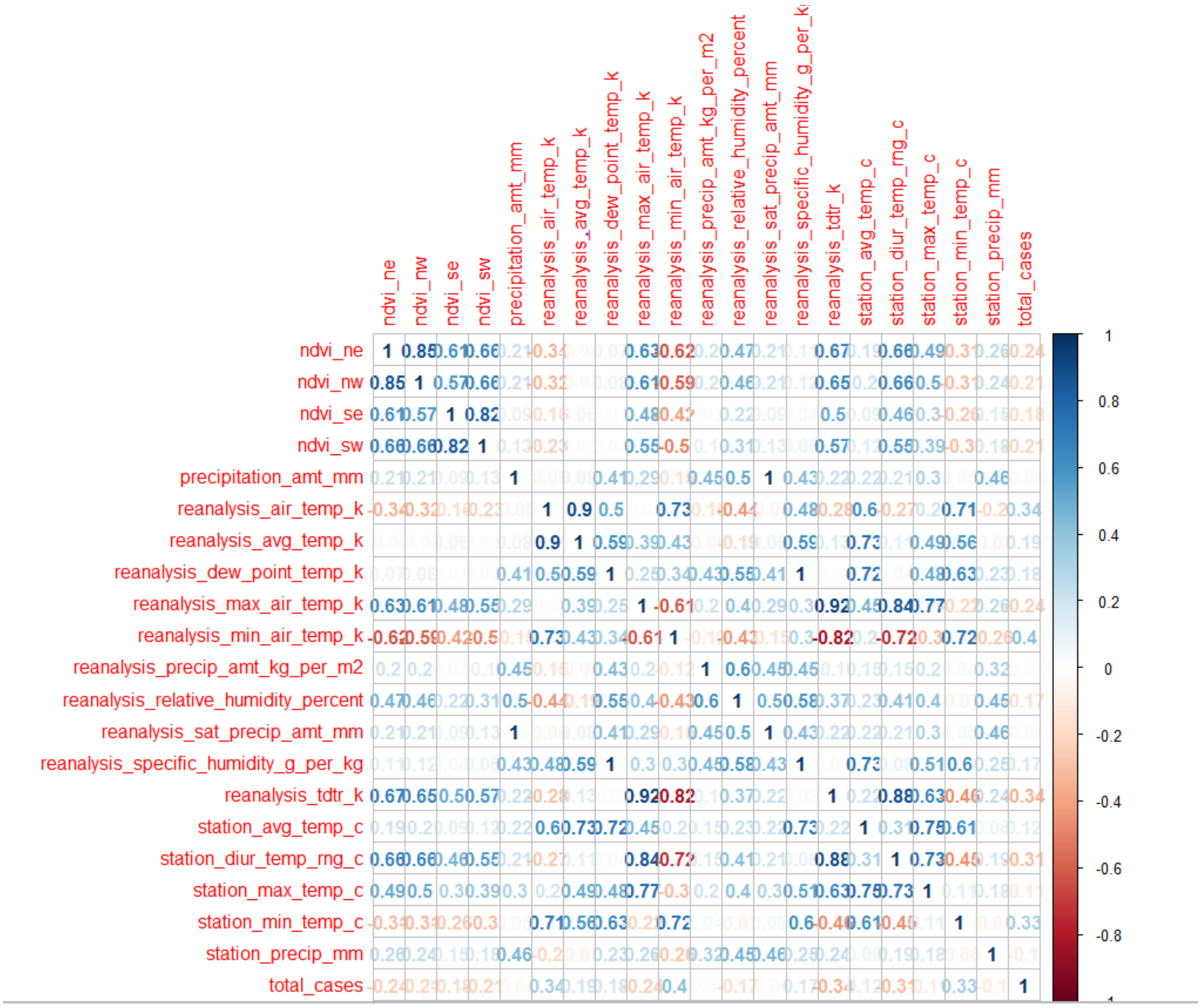
We also converted all Kelvin temperatures to Celsius to maintain the consistency between different temperature measures.

Total number of dengue cases vs Date

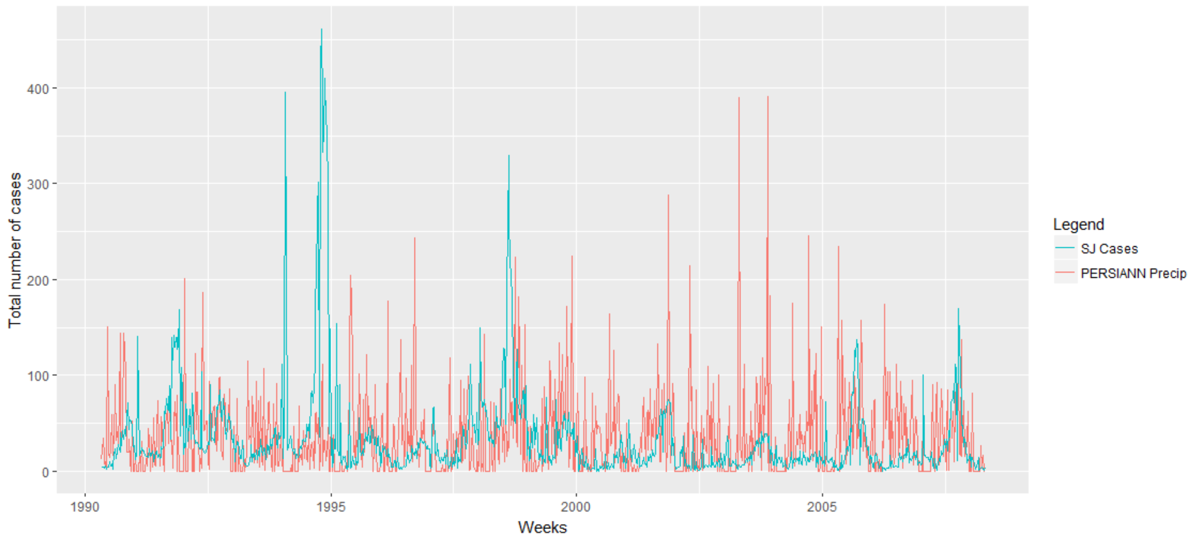


The above plot confirms the **strong seasonality of dengue transmission** and its punctuation by major outbreaks, as marked by sudden massive spikes in the plot. Their **spontaneity indicates that it might not be a good idea to predict these using time series analysis alone**. The difference in plot height between San Juan and Iquitos could be attributed to a big disparity in population size.

Correlation between number of cases and other attributes

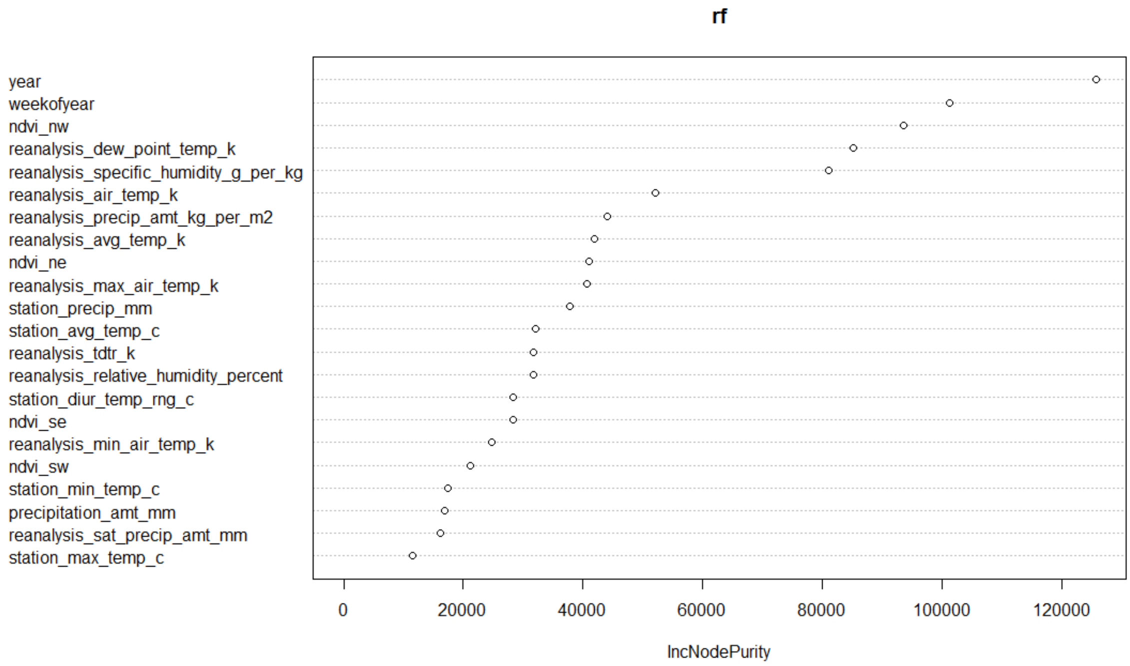


We see that none of the variables are significantly correlated with the total number of cases (although some of the temperature variables are correlated with each other, as expected). It may well be that our expected model is a combination of several of them, with no particular features having a strong impact on the outcome.



The plot exhibits phenomenon we may have come to expect from the literature -- the seasonality in precipitation somewhat mirrors that of dengue cases, but the lows in the dengue plot also fall on massive peaks in precipitation, which may be due to intense rainfall washing out mosquito breeding sites. Curiously, the two outbreaks between 1992 and 1995 were during periods with relatively low precipitation, suggesting the contribution of other variables.

We next plotted ranking of the variables in order of significance, by exploiting the algorithm used within random forests for deciding upon which variable to split.



Observations: weekofyear and year ranked as two of the most important.

We have deleted any features that are strongly correlated to others based on the corrplot.

We have removed the following attributes:

'city','reanalysis_tdtr_k','reanalysis_relative_humidity_percent','reanalysis_specific_humidity_g_per_kg','station_diur_temp_rng_c'

IV. Feature Engineering:

To make Machine learning algorithms more effective we use Feature engineering. It is nothing but transforming our features or come up with new features.

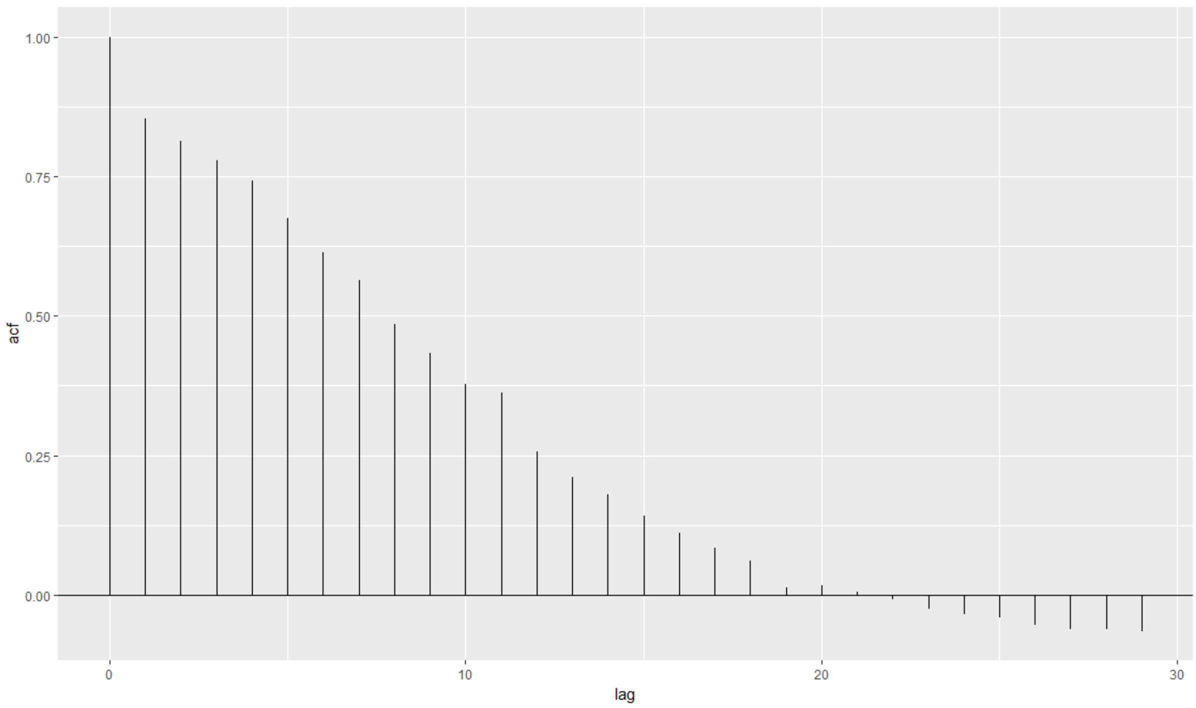
When you visually examine the total cases reported per week in the training files, can we verbally express that those cases are the result of the meteorological conditions recorded in previous weeks? What would be the relationship between the reported cases for a given week and the information provided for the same week.

What we can notice is that the infections reported for a specific week are not a direct cause-effect of the meteorological conditions that are observed on that specific week. The most likely cause of this has an incubation period of 4-7 days, which is that the infection is directly related to the conditions in previous weeks.

During data analysis, would it be appropriate to shift the data to one week back.

As seasonal time series plays an important role, we thought for each observation it would be a better idea to add variables on past time lags.

The below plot would give the correlation between the total number of cases at a given time with those of the past 15 weeks to the given time.



V. Model Training and Validation:

Performance Metrics Used: MAE, RMSE, R Squared and Predicted vs Actual plots

The mean absolute error (MAE) measures the closeness of forecasts or predictions to the actual outcomes. The mean absolute error is given by:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|.$$

$$AE = |e_i| = |y_i - \hat{y}_i|$$

$$Actual = y_i$$

$$Predicted = \hat{y}_i$$

The square root of the mean/average of the square of all of the error. It is generally used as an error metric for numerical predictions.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

<u>K-NN</u>

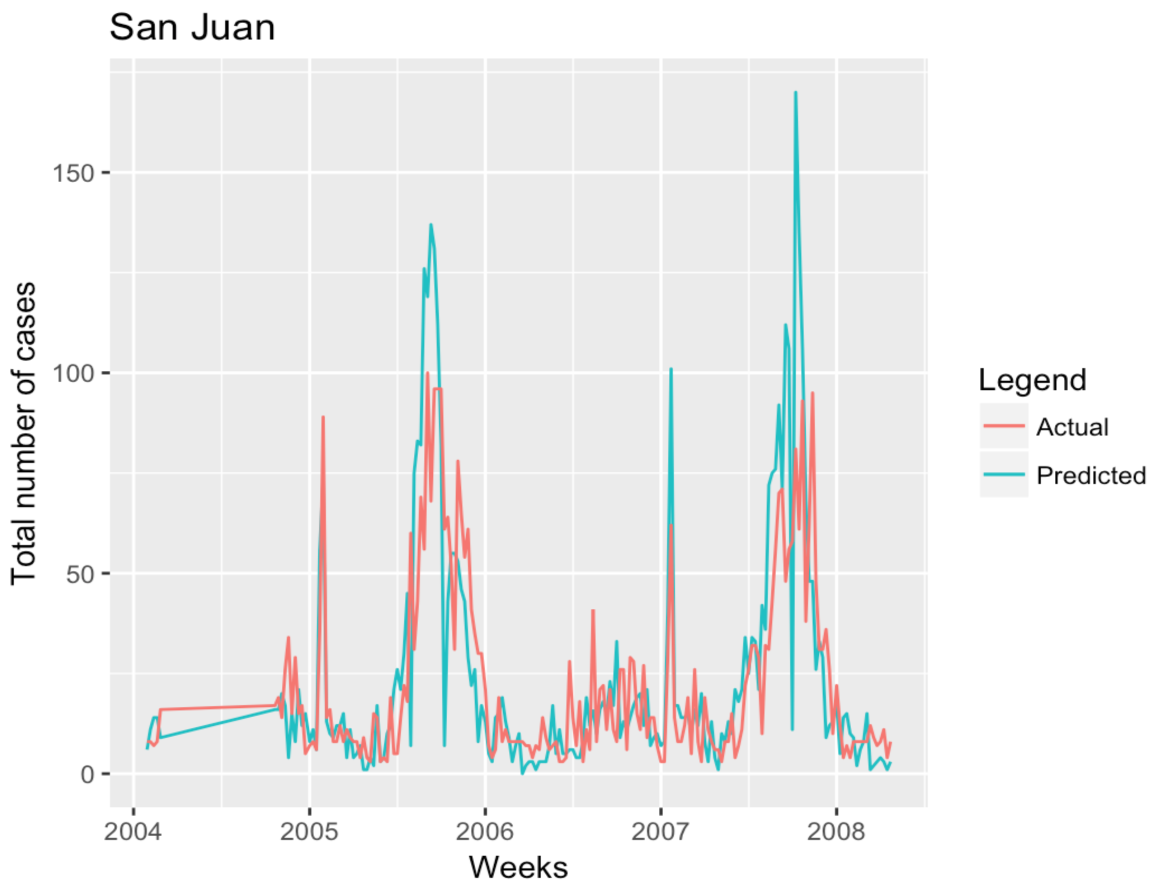
Parameters:

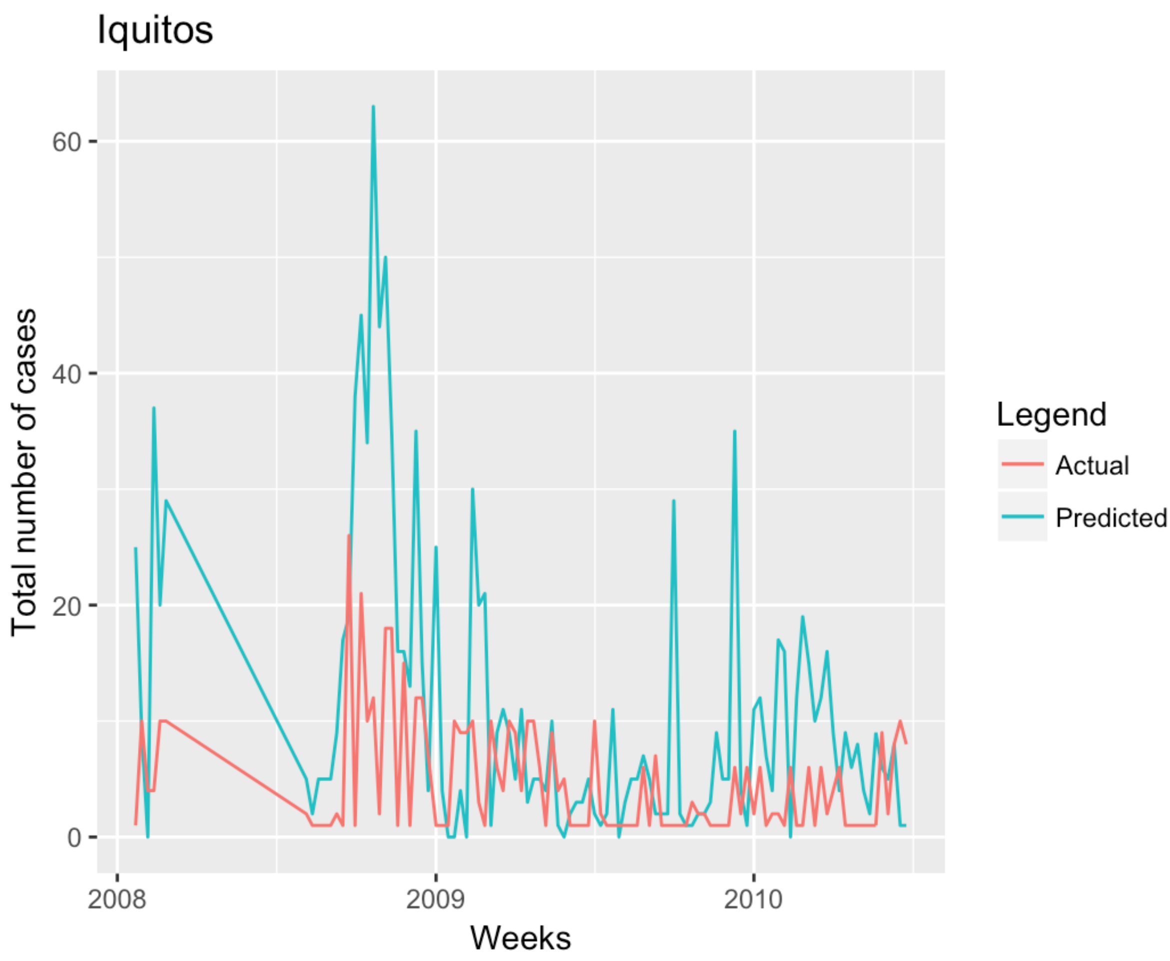
K: It states the number of neighbours considered.

L: It states minimum number of votes for definite decision

Classifier	**K**	**L**	**MAE**	**RMSE**
K-NN	10	1	Sj: 12.13298 Iq: 8.580952	19.3004 12.44531
K-NN	12	1	Sj: 21.84574 Iq: 8.666667	37.52722 13.01208
K-NN	20	2	Sj: 22.18617 Iq: 8.666667	38.08983 13.01208
K-NN	60	5	Sj: 22.18617 Iq: 8.666667	38.08983 13.01208
K-NN	80	10	Sj: 22.18617 Iq: 8.666667	38.08983 13.01208

Predicted vs Actual cases plot:





SVM

Parameters:

Gamma: parameter needed for all kernels except linear.

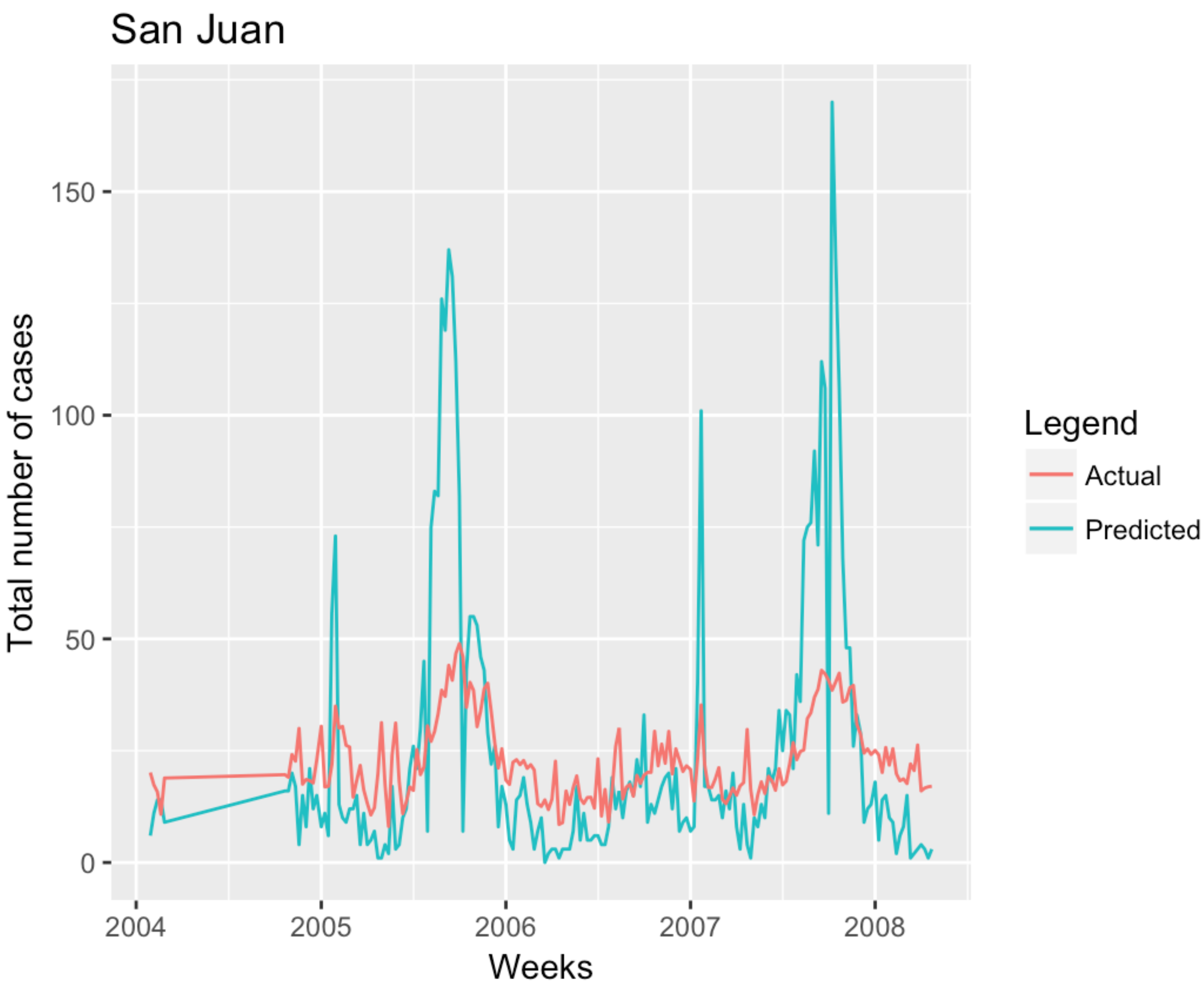
Epsilon: epsilon in the insensitive-loss function

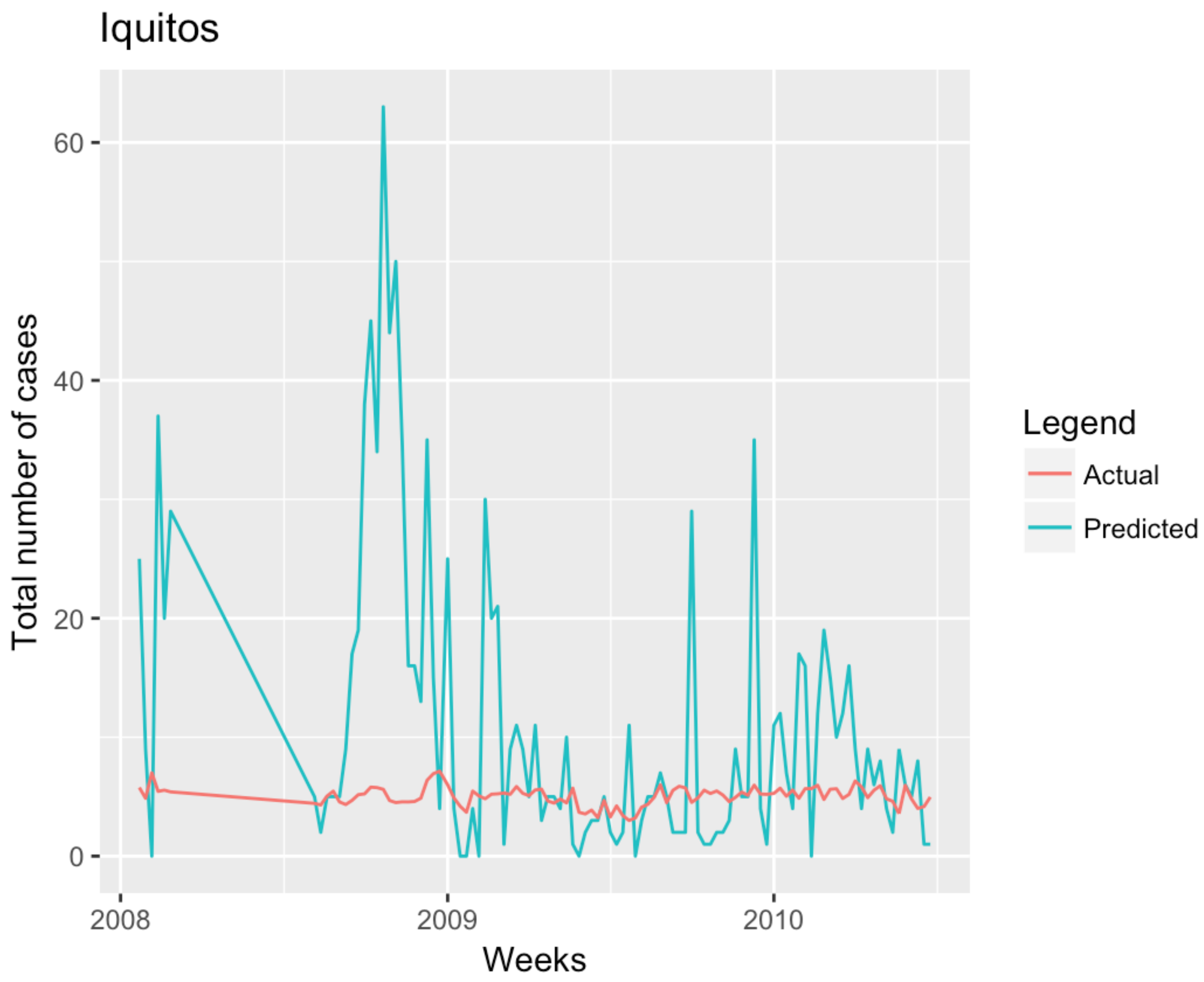
Cost: cost of constraints violation

Classifier	gamma	epsilon	cost	MAE	RMSE
SVM	0.08	0.1	0.05	Sj: 16.82847 Iq: 8.214406	26.90979 13.70953
SVM	0.08	0.1	0.09	Sj: 16.34536 Iq: 8.133187	25.86538 13.5947
SVM	0.05	0.2	0.05	Sj: 16.42978 Iq: 8.001587	24.4679 13.3189
SVM	0.056	0.3	0.095	Sj: 19.63812 Iq: 7.956617	26.03124 13.04673

SVM	0.1	0.8	0.06	Sj: 37.5701 Iq: 8.545063	40.19254 12.46896

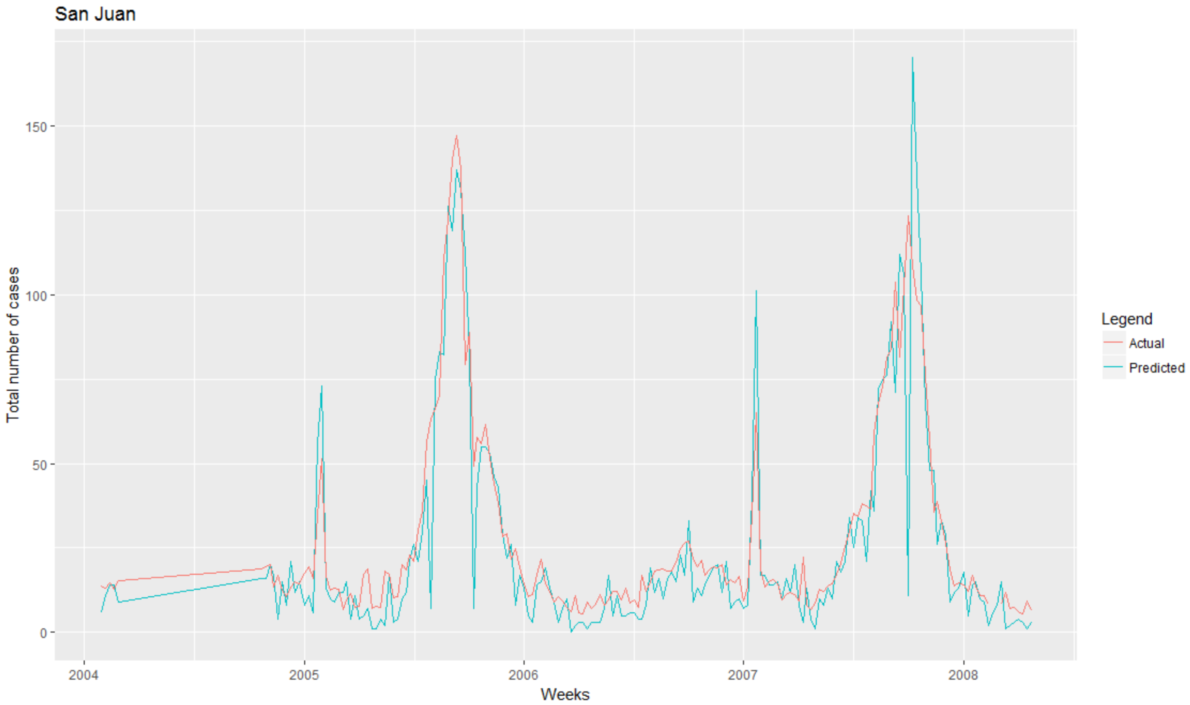
Predicted vs Actual cases plot:

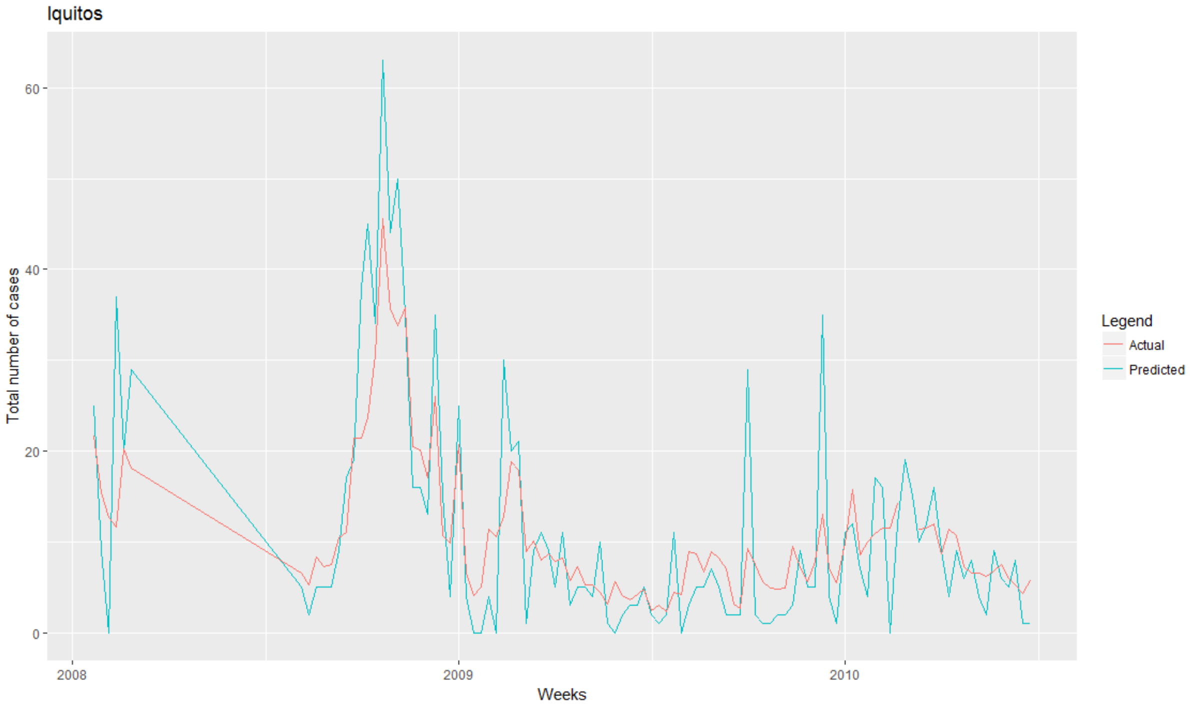




Classifier	ntress	mtries	maxdepth	MAE	RMSE
Random Forest	1000	3	4	Sj: 16.027 Iq: 4.605	30.4456 8.914
Random Forest	1000	4	4	Sj: 15.258 Iq: 4.569	29.86 8.9308
Random Forest	1000	5	4	Sj: 14.758 Iq: 4.537	29.459 8.938
Random Forest	1000	5	5	Sj: 14.00 Iq: 2.286	28.66 8.88
Random Forest	1000	10	5	Sj: 13.472 Iq: 4.546	28.73 9.11
Random Forest	1000	10	10	Sj: 13.2106 Iq: 4.5516	28.66 9.101

Predicted vs Actual cases plot:

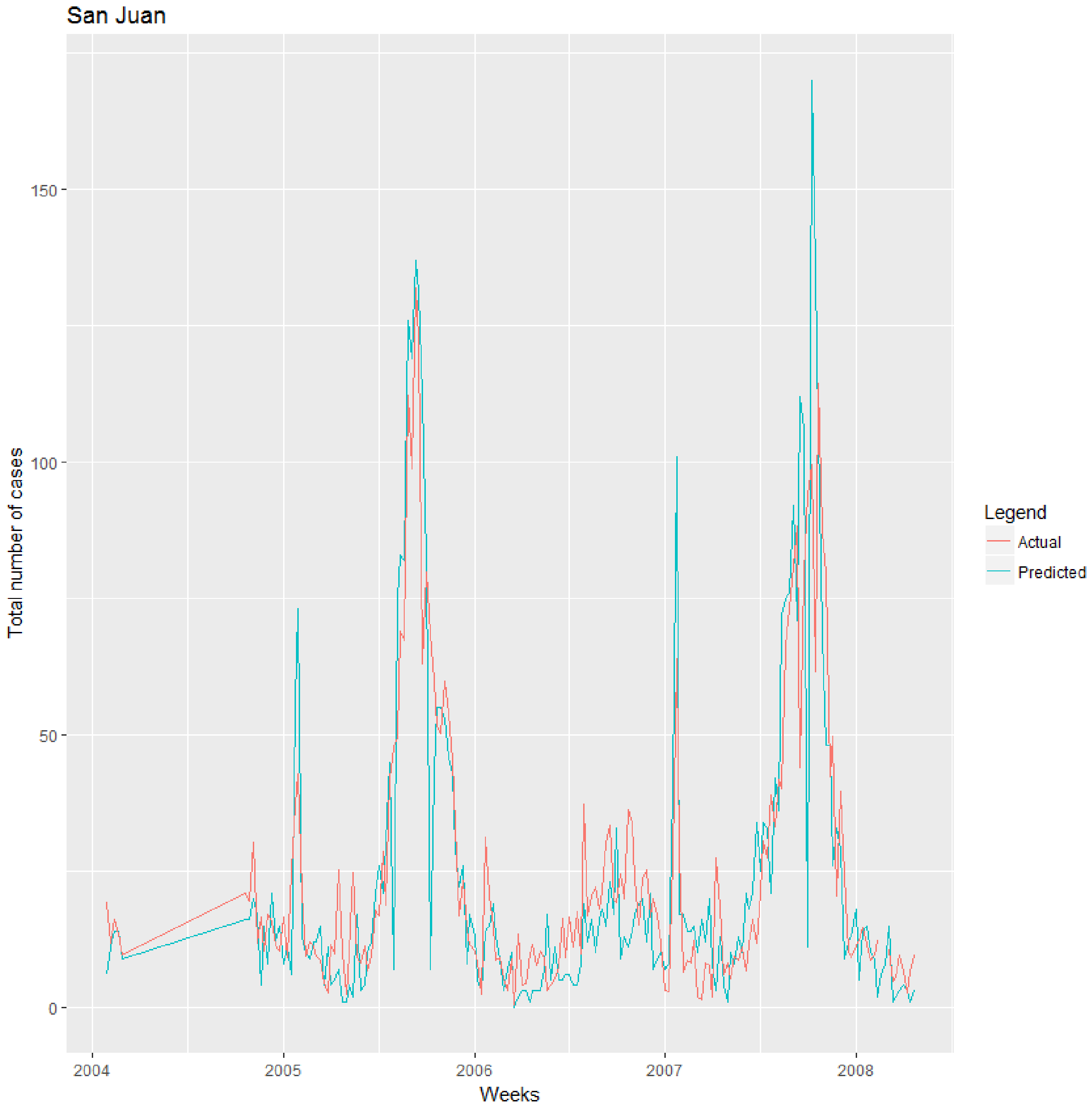


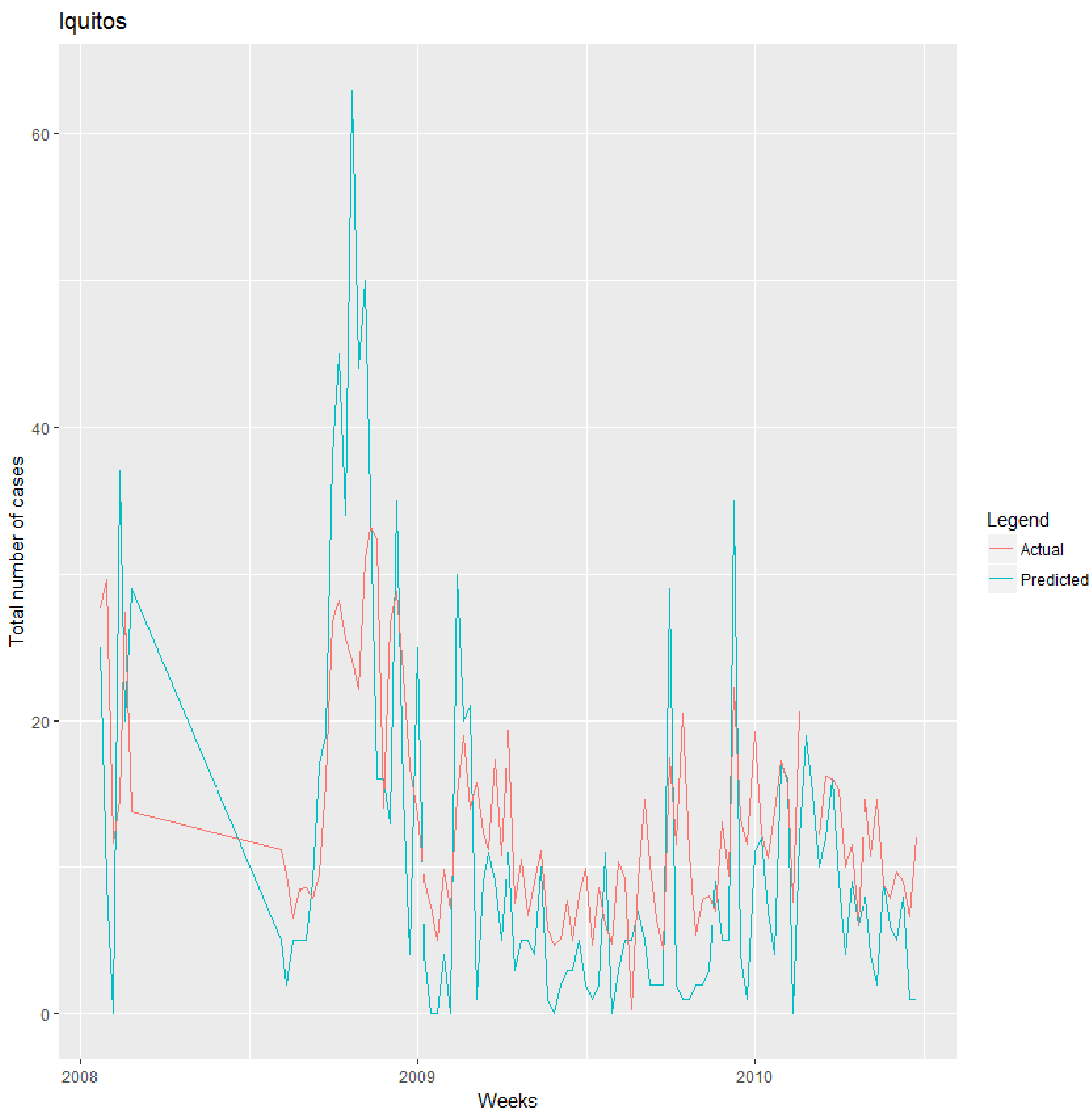


Classifier	**Activation function**	**Hidden layers**	**MAE**	**RMSE**
Deep Learning	TanhWithDropout	5 c(50,50,50,50,50)	sj:23.41876 iq:4.809866	sj:43.5228 iq:9.108329
Deep Learning	TanhWithDropout	4 c(50,50,50,50)	sj: 21.31624 iq: 4.426676	sj: 41.32863 iq 8.767354

Deep Learning	TanhWithDropout	7 c(50,50, 40,50,70, 40,50)	sj: 21.87453 iq: 5.164498	sj: 44.99929 iq: 9.332911
Deep Learning	Rectifier	7 c(50,50, 40,50,70, 40,50)	sj:12.8167 iq:4.558575	sj:25.57317 iq:7.061114
Deep Learning	Rectifier	4 c(50,50,50,50)	sj: 13.88554 iq: 5.524445	sj: 25.24392 iq: 7.079772
Deep Learning	Rectifier	5 c(50,50,50,50,50)	sj: 13.45377 iq: 5.515761	sj: 24.40207 iq: 9.062706

Predicted vs Actual cases Plot:

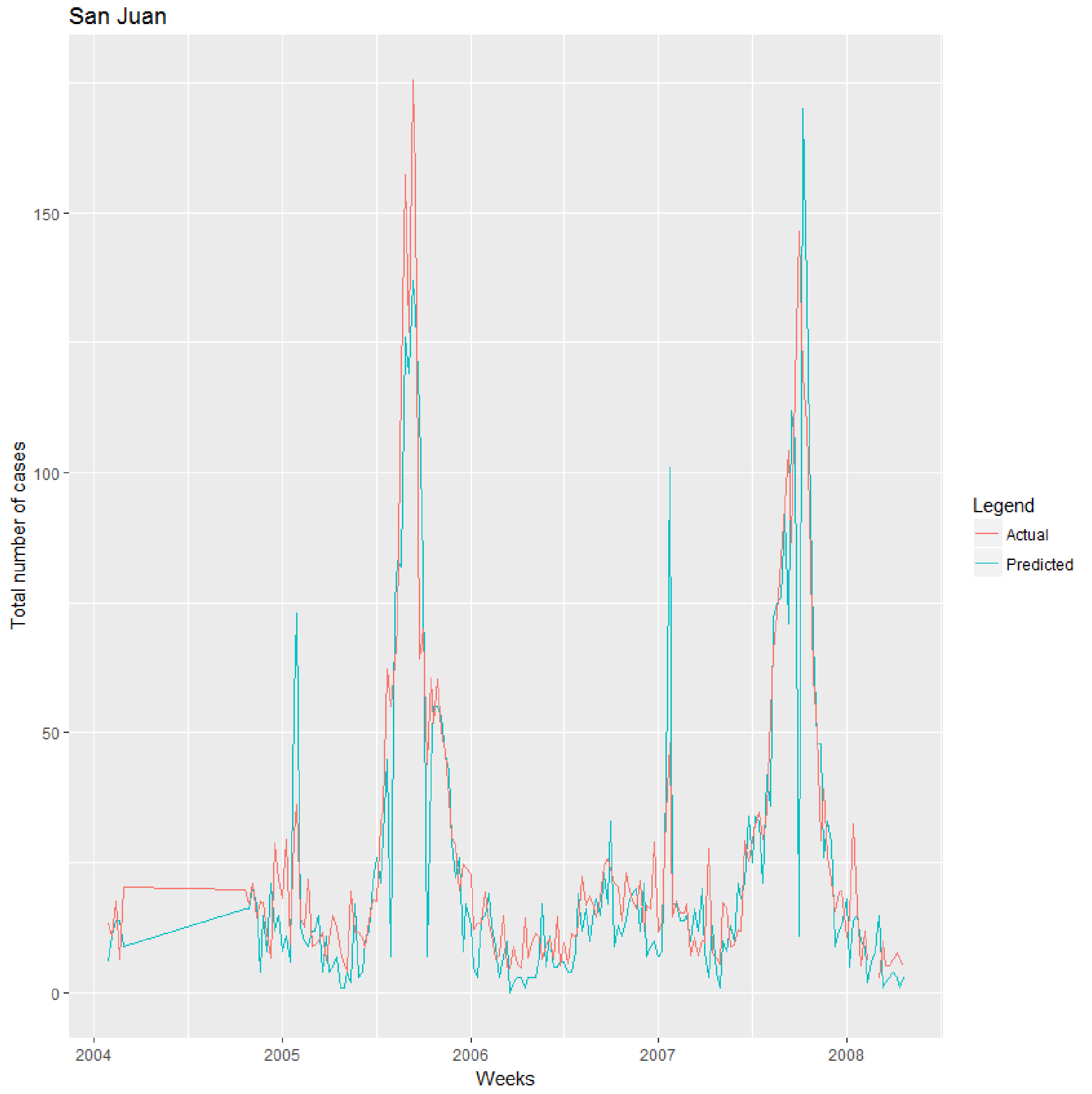

San Juan
Total number of cases
150
100
50
0
2004
2005
2006
2007
2008
Weeks
Legend
Actual
Predicted

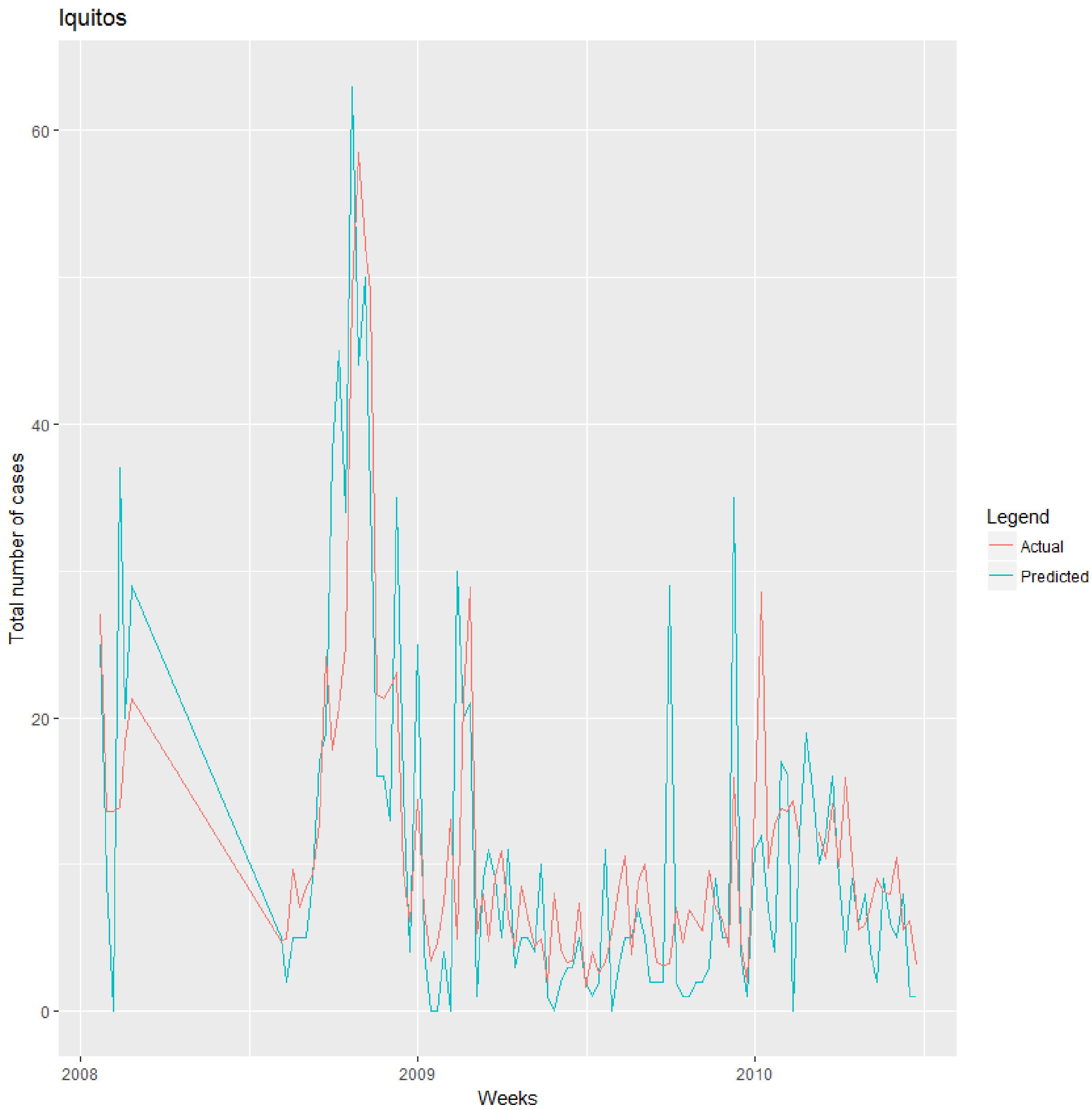


Classifier	ntress	Stopping rounds	maxdepth	MAE	RMSE	R^2
Gradient Boosting	1000	10	4	sj: 8.314 iq: 2.337	sj: 13.979 iq: 4.111	sj: 0.9354 iq: 0.835
Gradient Boosting	1000	15	5	sj: 7.623992 iq: 2.207122	sj: 13.25064 iq: 4.011511	sj: 0.9419864 iq: 0.8430503
Gradient Boosting	1000	5	8	sj: 5.129772 iq: 1.648242	sj: 10.86593 iq: 3.535703	sj: 0.9609887 iq: 0.878074
Gradient Boosting	800	12	20	sj: 2.951952 iq: 1.114673	sj: 10.73635 iq: 3.607621	sj: 0.9619136

						iq: 0.873 0635
Gradient Boosting	1200	15	25	sj: 1.854691 iq:0.7549577	sj: 7.823227 iq: 2.853748	sj: 0.9797 778 iq: 0.920571 6
Gradient Boosting	1500	20	30	sj:1.4092 iq: 0.5992183	sj: 6.284714 iq: 2.421354	sj: 0.9869 495 iq: 0.942817 8

Predicted vs Actual Plot:





VI. Conclusion:

Among all the classifiers that we used, **Gradient Boosting** gave us the best results. This is because Gradient boosting build trees sequentially one at a time, where each new tree helps to correct the previously trained tree errors. While each new tree added, the model becomes more expressive.

We used four parameters mainly and they are the number of trees, depth of trees, stopping rounds and learning rate. We tested by modifying the parameter values each time. While increasing trees and depth of the trees, we got better accuracy.

We also made our submission on test data to Driven data competition and we stand at 11th position in the competition. Here is the snapshot of our submission and rank.

DengAI: Predicting Disease Spread

HOSTED BY DRIVENDATA

Submissions

BEST SCORE	CURRENT RANK	# COMPETITORS	SUBS. TODAY
22.3726	11	712	3 / 3

EVALUATION METRIC

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$$

In future we will try to improve this score and we aim for winning the competition.

END OF REPORT