Prediction of West Nile virus in Chicago

LAL BABU SAH

MAY 30, 2020

Introduction:

In this project separate train, test, weather, and spray dataset are given. And I have to predict whether or not ‘West Nile virus’ in Chicago is present, for a given time, location, and species using given dataset. Here, I will use the Classification machine learning algorithm for the prediction of ‘West Nile virus’.

Data Acquisition:

Read the train, test, weather, and spray dataset using pandas.

Data Wrangling:

First merged ‘train’ dataset with ‘weather’ dataset along the common column (Primary Key) ‘Date’. Then I have matched the ‘Date’ and ‘location’ of the GIS dataset with the merged dataset and added the new column(spray) in the merged dataset (added spray is equal to 1 if matched else 0).

After that, I have dropped some columns which are not important for prediction, like from (‘Block’, ‘Address’, 'Address Number and Street') these all contains address, dropped ‘Address’, 'Address Number and Street' but not ‘Block’.

Column (‘species’) contain different types of species name which are the object type, I have converted these in numerical data. Some columns contain the numerical data, but it is the object type, converted these into ‘int’, or ‘float’ type. Some columns ('StnPressure', 'SnowFall') contain ‘M’ which are missing data, replaced it with ‘0’.

Data Exploration:

Calculated the Pearson Correlation Coefficient and the p-value of different columns and dropped the columns which are not important for prediction on the basis of correlation and p-value. Checked the outliers of different columns, to remove the outliers if present.

Model Development:

After data exploration normalized the dataset and then split into the train and test dataset. Here I have used classification machine learning algorithm because target ('WnvPresent’) is categorical (‘Yes’, ‘No’). After that, preprocess the test dataset and predicted the 'WnvPresent’.

Data Visualization using Tableau:

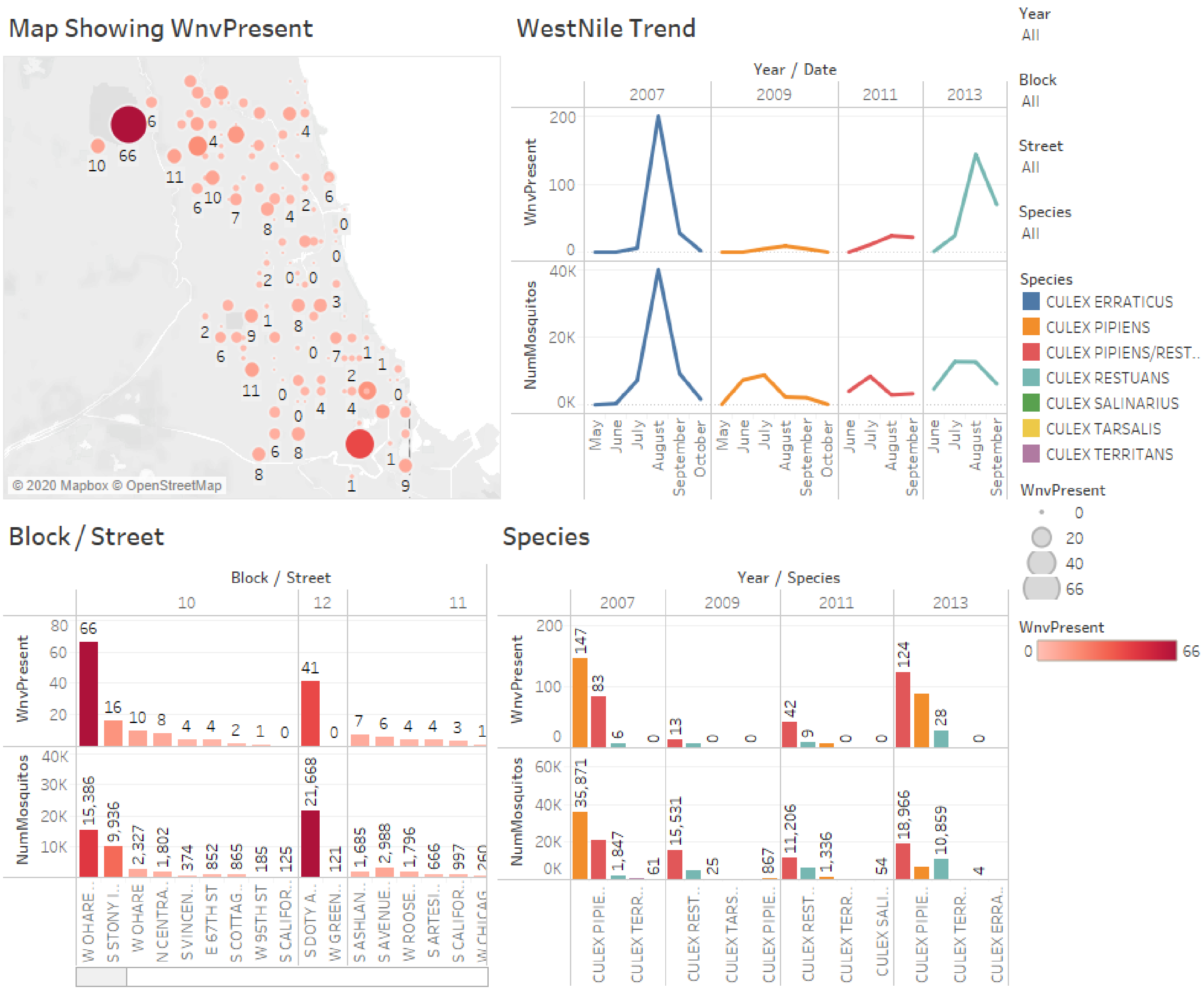
In tableau uploaded the merged train dataset for the dashboard preparation.

Using geographical coordinates plotted map, which is showing in which location ‘West Nile virus’ is present more or less for different years.

Using the columns ‘Date’ and 'WnvPresent’ plotted year wise trend, which shows in which month and year ‘West Nile virus’ is present more or less.

Using columns ‘Block’, ‘Street’, 'NumMosquitos', and 'WnvPresent’ plotted bar graph, which shows the number of 'NumMosquitos' and 'WnvPresent’ for a given ‘Block’ and ‘Street’ for different years.

Using columns ‘Species’ and 'WnvPresent’ plotted graph, which shows the number of 'WnvPresent’ for different species for different years.



Map Showing WnvPresent
66
10
11
© 2020 Mapbox © OpenStreetMap
WestNile Trend
Year / Date
2007
2009
2011
2013
WnvPresent
200
100
0
NumMosquitos
40K
20K
0K
May
June
July
August
September
October
Year
All
Block
All
Street
All
Species
All
Species
CULEX ERRATICUS
CULEX PIPIENS
CULEX PIPIENS/REST..
CULEX RESTUANS
CULEX SALINARIUS
CULEX TARSALIS
CULEX TERRITANS
WnvPresent
0
20
40
66
WnvPresent
0
66
Block / Street
Block / Street
10
12
11
WnvPresent
80
60
40
20
66
16
10
8
4
4
2
1
0
41
0
7
6
4
4
3
1
NumMosquitos
40K
30K
20K
10K
15,386
9,936
2,327
1,802
374
852
865
185
125
21,668
121
1,685
2,988
1,796
666
997
260
W OHARE ..
S STONY I..
W OHARE
N CENTRA..
S VINCEN..
E 67TH ST
S COTTAG..
W 95TH ST
S CALIFOR..
S DOTY A..
W GREEN..
S ASHLAN..
S AVENUE..
W ROOSE..
S ARTESI..
S CALIFOR..
W CHICAG..
Species
Year / Species
2007
2009
2011
2013
WnvPresent
200
100
0
147
83
9
0
13
0
0
42
9
0
0
124
28
0
NumMosquitos
60K
40K
20K
0K
35,871
1,847
61
15,531
25
867
11,206
1,336
54
18,966
10,859
4
CULEX PIPIE..
CULEX TERR..
CULEX REST..
CULEX TARS..
CULEX PIPIE..
CULEX REST..
CULEX TERR..
CULEX SALI..
CULEX PIPIE..
CULEX TERR..
CULEX ERRA..