Title: Analysis of the Detroit Blight Occurrences using Neural Net Predictors

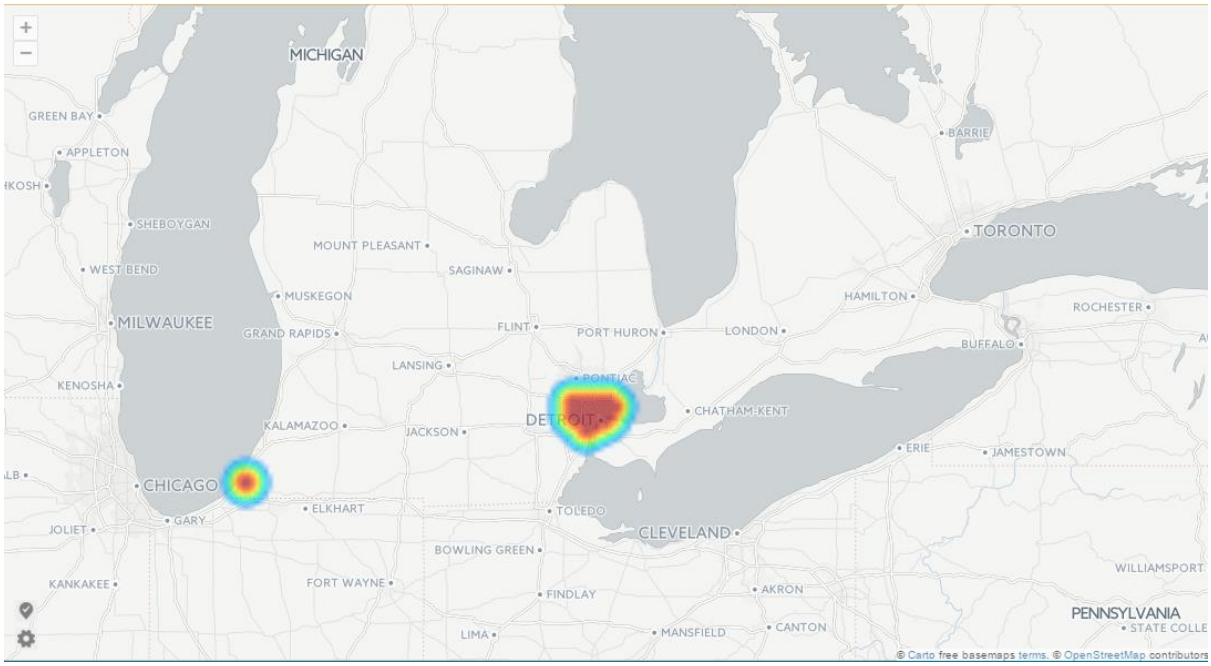
Overview

The analysis was primarily done using R-Caret library which provides re-useable components for predictive analysis. Data preparation was done using SQL fucntions in Mysql database

From the analysis that was done, it was found out that there are key features that can be used to predict occurrence of blight. This includes fines that were levelled to occupants of buildings which could determine the nature of violation that had occurred in the building.

I also proceeded to perform Geographic clustering of violations to determine which buildings were affected and there were Clusters and from this it seems key areas that stand out in terms of blight incidences occurrence include the areas of Dearborn Heights, Highland Park hence areas around this region were likely to have incidences of blight.

There were also outliers in the data as can be seen with the cases reported in area of Chicago in the below visual.



Methodology

The methodology used include:

- loading the training data and structuring the schema of the data in Mysql Community Edition

- Data preparation was done in Mysql ; please see the identification and data cleaning steps below

- analysis of the data to determine missing input values using KNN algorithm. Any missing data was then estimated and filled to normalize the data

- The data also had numeric and string variables which needed to be converted into categorical factors with levels.

- Data was then split into 5-folds using Cross-validation technics to randomise then partition into 5 folds of which 1 fold would be used as test data and rest training sets.

- Once training and test sets were available I proceeded to perform Feature selection to identify categories that would be most useable in Predictive analysis at which I arrived at 5 predictor classes - Accuracy was used to select the optimal model using the largest value.

Identification of blighted buildings:

a. Obtain all 311 complaint tickets with the following issue types i.e. 'Running Water in a Home or Building', 'Graffiti', 'Graffiti Abatement (internal use only, public issue)'. This would ideally be indicative of a building.

b. Obtain records from blight violations with Violation descriptions indicative of existence of a building i.e. descriptions like building, dwelling. Then proceed to obtain from Crime file any crimes reported that could have happened in a building – indicative of existence of a building e.g. burglary, arson

c. Obtain the demolition permits that have descriptions that mention buildings. These would form the initial list of identified buildings.

```
3
4 INSERT INTO ListOfBuildingsIdentified(city, `GeoMap`,311Complaints)
5
6 SELECT
7     distinct
8
9     `detroit_311complainttickets`.`city`, concat(`detroit_311complainttickets`.`lat`,',', `detroit_311complainttickets`.`lng`) as GeoMap,
10    '311Complaints'
11
12 FROM
13    `dataimports`.`detroit_311complainttickets`
14 WHERE
15    issue_type IN ('Running Water in a Home or Building' , 'Graffiti',
16        'Graffiti Abatement (internal use only, public issue)')
17
18
19 union
20
21 select distinct MailingAddress,GeoMap_1, 'BlightViolations' from detroit_blightviolations where
22  ( ViolDescription like '%building%'  or  ViolDescription like '%dwelling%' or
23  ViolDescription like '%Rodent%' or ViolDescription like '%property%')
24
25  union
26
27  select  Address, GeoMap_1, 'Crime' from   detroit_crime where Category in (
28 'DAMAGE TO PROPERTY'
29 'BURGLARY',
30 'ARSON',
31 'KIDNAPING',
32 'STOLEN PROPERTY',
33 'OTHER BURGLARY',
34 'CONSPIRACY BY COMPUTER'
35 ) and Offensedescription NOT LIKE '%MOTOR VEHICLE%' and Offensedescription NOT LIKE '%TRAFFIC VIOLATION%'
36 and Offensedescription NOT LIKE '%VEHICLE%';
37
38
```

d. Obtain and setup 20m cordon radius from the geo-coordinates available in the demolitionpermits. The demolition permits indicate the houses that have incidences of blight.

```
4 -- --
5 UPDATE detroit_demolitionpermits set latX1=(lat + (20 * 0.0000089)), longY1=(`long` + ((20 * 0.0000089)/ COS(lat * 0.018))),
6  latX2=(lat - (20 * 0.0000089)) , longY2= (`long` - ((20 * 0.0000089)/ COS(lat * 0.018)));
7 --
8
9
```

e. The last step was to match the records obtained from 311complaints tickets, blight violations and crimes file and check if they fall within the 20m cordon radii segments defined in (d) above. If a record is within the cordon then it is marked as blighted. Procedure below:

```
drop procedure procCheckBlight;

DELIMITER $$
CREATE PROCEDURE procCheckBlight()
DETERMINISTIC
BEGIN
    DECLARE i int DEFAULT 0;

    DECLARE pcity varchar(255);
    DECLARE pGeoMap varchar(255);
    DECLARE p311Complaints varchar(255);
    DECLARE platx double;
    DECLARE plongy double;
    declare done int;


       declare cur1 cursor for
     select city, `GeoMap`,311Complaints,latx,longy
     from ListOfBuildingsIdentified where Blighted <> 1;

  declare continue handler for not found set done=1;

     set done = 0;
     open cur1;
        igmLoop: loop
           fetch cur1 into pcity,pGeoMap,p311Complaints,platx,plongy;
           if done = 1 then leave igmLoop; end if;

           IF  EXISTS(select 1 from detroit_demolitionpermits where latX1 <=platx and latX2 <=platx and longY1 >= plongy and longY2 <= plongy ) THEN
               update  dataimports.ListOfBuildingsIdentified set Blighted=1 where latx=platx and longy = plongy;
               insert into IterationReport(city,GeoMap,`311Complaints`,latx,longy,Blighted) values (pcity,pGeoMap,p311Complaints,platx,plongy,1);
               #SELECT 'Blighted Building ....';
           ELSE
              insert into IterationExceptionReport(city,GeoMap,`311Complaints`,latx,longy,Blighted)  values (pcity,pGeoMap,p311Complaints,platx,plongy,0);
                  #SELECT 'Non-Blighted Building ....';
           END IF;
       end loop igmLoop;

     close cur1;

END$$
DELIMITER ;
```

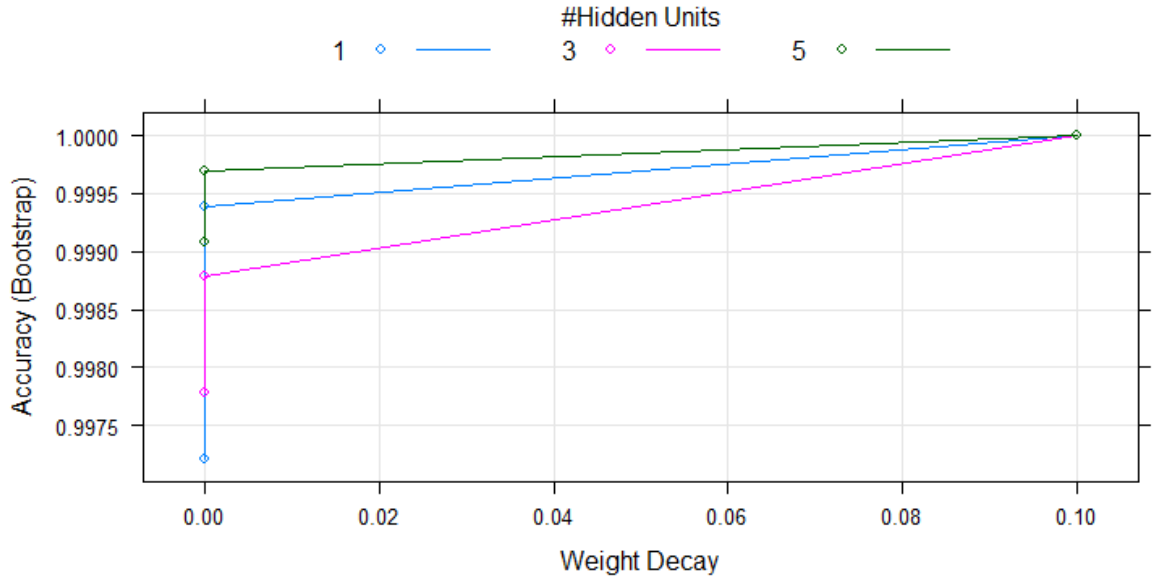
Building the Prediction Model:

The predictor class included the following variables:

- outcome,
- PaymentStatus,
- ViolationCategory,
- Disposition,
- AgencyName,
- Building Type Usage

Using the neural network library in Caret I was able to check and find out the relevance of the above predictors.

The attached snapshot is a capture of the accuracy of the Neural Net model training results. I believe the method is effective and efficient compared to two other algorithms tried out i.e. Gradient Boost Method and Generalised Linear Model - which required more compute time to machine learn the underlying patterns.



The results seem also promising for the Neural Net approach as per below:

Neural Network

198806 samples

5 predictor

2 classes: '0', '1'

No pre-processing

Resampling: Bootstrapped (25 reps)

Summary of sample sizes: 198806, 198806, 198806, 198806, 198806, 198806, ...

Resampling results across tuning parameters:

size	decay	Accuracy	Kappa
1	0e+00	0.9972133	0.6371300
1	1e-04	0.9993873	0.9198914
1	1e-01	1.0000000	1.0000000
3	0e+00	0.9977784	0.7123525
3	1e-04	0.9987879	0.8400000
3	1e-01	1.0000000	1.0000000
5	0e+00	0.9990741	0.8800000

5 1e-04 0.9996864 0.9600000

5 1e-01 1.0000000 1.0000000

Hopefully, this method can be used for Predictive analysis to determine beforehand where occurrences of blight are likely to be. There is also a wealth of data in the narrative fields e.g. violation description, offenses description, complaint description from the 311 logs.

Since the data was captured in natural language format, one could leverage Big Data algorithms such as MapReduce and Natural Language Processing to determine frequency of text/terms occurrence, context of phrases and sentiment. This would be used as a counter-check to augment the Predictive algorithm built above to predictively machine learn future occurrences of blight in buildings.