Detroit Blight Analysis

CAP STONE

Bala Subrahmanyam | Data Science at Scale, Coursera.org, No Restrictions | January 1, 2017

Get Started Right Away

This project is to analyze Detroit blight data to find appropriate model to predict future demolish judgments. The issue is due to various factors like crime, environment maintenance issues, neglecting properties due to people moving out of the town and poverty. We have different dimension data associate with a location like citizens reported issues and agency recorded incidents along with crime data.

Figure 1 : Data Dimensions

Data Representation:

Visualization to eliminate outliers, used CARTO and Microsoft excel power view to prepare data out of 311 calls and crime data sets as layers. Merged blight violation data set on those units to prepare data.
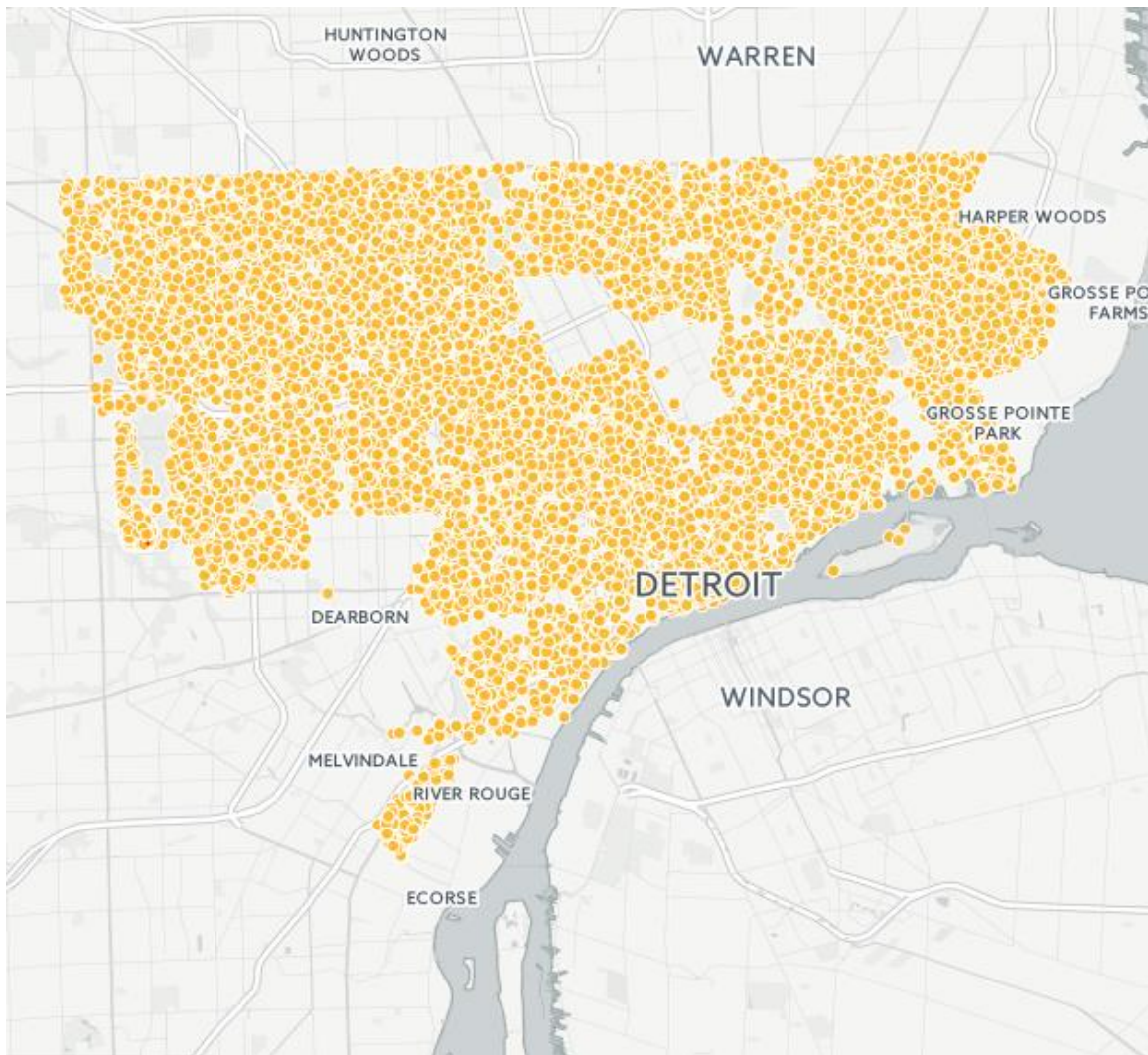
https://tubatibs.carto.com/builder/cf461ff8-c349-11e6-85e3-0ecd1babdde5/embed

Figure 2: Data visualization and analysis

Understanding Data:

detroit-blight-violations.csv : Each record is a blight violation incident. Contains data associated to location observed and raised by different agencies.

detroit-demolition-permits.tsv: Each record represents a permit for a demolition. Contains fees and associated financial parties with dates.

detroit-311.csv: Each record represents a 311 call, typically a complaint. Contains citizen reported issues.

detroit-crime.csv: Each record represents a criminal incident. Contains location and details about criminal activity.
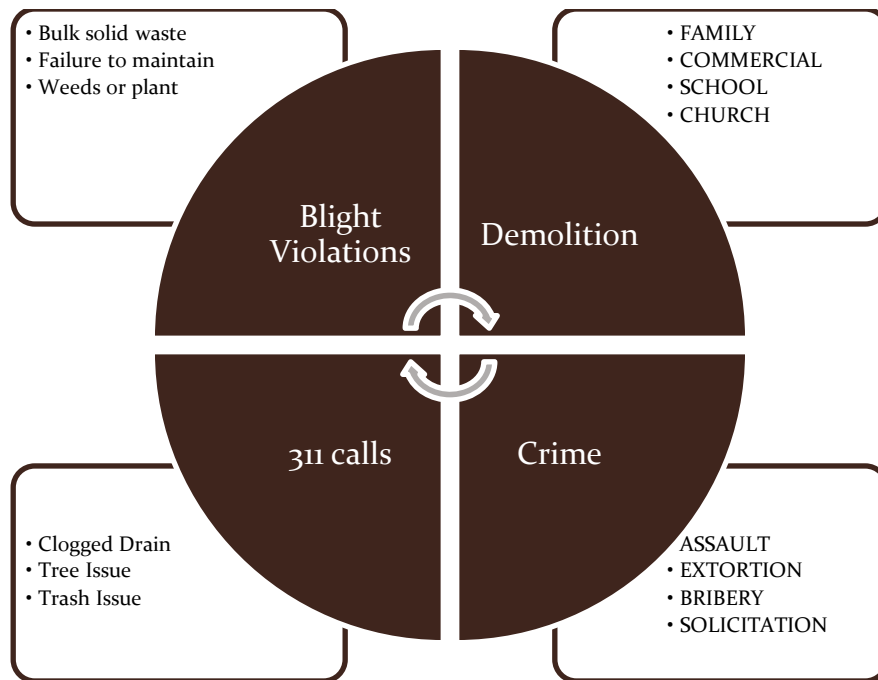
Figure 3: Data matrix with few items

Steps to capture data:

1. Filter out noisy data in all four files.
2. Extract latitude and longitude incase if it is part of address.
3. Crime data inner join with Blight violations aggregated crime count associated with location.
4. Left outer join with 311 calls data with count of issue type and sum of ratings.
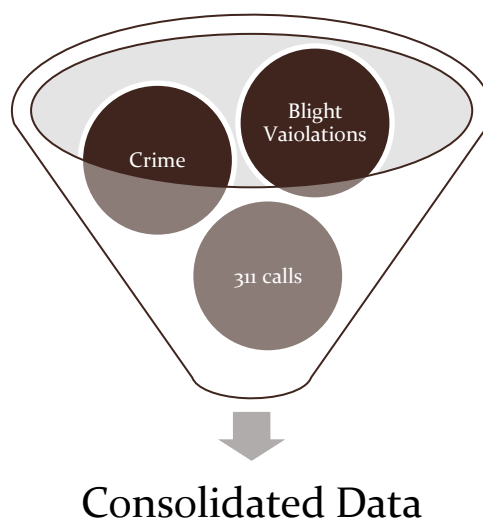5. Validated consolidated data for any NaN or special characters.

Figure 4: Data funnel representation

Training Dataset:

Mapped demolition permits records on above consolidated data set to derive demolished buildings. Other records will be treated as still in-use. Explored different models to validate predictions.

Evaluation of Advanced Predictive Data Models:

Data cleansing took long time to identify the important attributes with respect to available data. Using R programming to evaluate models. Explored predictive models as follows

- Linear Regression model:
- SVM
- Random Forest Model
- XGBoost

Formula explored to build models

formula <- BLIGHTED ~ ViolDescription + PaymentStatus + count_issue_type + sum_rating + COUNT_OFFENSE_CATEGORY

Calculated Root-Mean-Square-Error (RMSE) to evaluate error rate for each model.

Not sure about the values filled in place of string may impact accuracy, but may need repeat data and features evaluations for change in error rate.

Optimizing features to improve accuracy:

Optimized above formula with judgment amount to see any difference in error rate, but not observed much difference.

If I have enough time, I would like to

1. Evaluate how it works with derived sentiment values based on 311 call description and crime offence description.
2. Dig data deep into owner address outside Detroit and outside USA.
3. Would like to analyze housing, Job market and Schools.
4. Adding more dimensions data will help predict accurately.

BIBLIOGRAPHY

https://www.theguardian.com/money/2014/sep/28/detroit-demolish-ruins-capitalists-abandoned-buildings-plan
https://en.wikipedia.org/wiki/Decimal_degrees