This repository Search Pull requests Issues Gist

uwescience / **datasci_course_materials**

Watch 369 Star 678 Fork 2,289

Code Issues 4 Pull requests 19 Wiki Pulse Graphs

Branch: master **datasci_course_materials** / capstone / blight / **blightfight.md** Find file Copy path

billhowe Added readings and modified assignment description. 13eb586 a day ago

1 contributor

```
130 lines (66 sloc)   11.1 KB
```

Raw Blame History

Blight Fight Capstone Project

"Urban blight refers to the deterioration and decay of buildings and older areas of large cities, due to neglect, crime, or lack of economic support. This is a typical sight in most US cities, and in many cities throughout the world. As a city gets older, some buildings or properties are not maintained and become run-down, abandoned or condemned." [1]

Many cities are actively trying to predict which properties are likely to become officially classified as blighted ahead of time. Predicting which properties are likely to be classified as blight can help cities take preventative action: a targeted demolition or renovation can prevent the spread of urban blight and facilitate economic revitalization of often distressed areas. The city of Detroit is one city interested in taking these measures, and city planners are actively pursuing blight prediction models. The city of Detroit also, like many cities, maintains an open data portal where data is published, often in live streams, to empower outside analysts to assist in problem solving and application construction.

In this assignment, you will work with real data to help urban planners predict blight. This is a real-world problem: the data will not be perfectly clean, the questions will not be perfectly unambiguous, and your results will not be perfectly reliable. But your work has the potential to help improve one city's economic future, and perhaps lead to a number of other cities following suit. And, we think you'll learn a lot!

Week 1: Background and Preparation

Get the Data

The data for the project is available in the github repository.

Additional data sets may be used in step 5. Up-to-date data can be downloaded or accessed via API from: data.detroitmi.gov

There are four files:

- `detroit-blight-violations.csv` : Each record is a blight violation incident.
- `detroit-demolition-permits.tsv` : Each record represents a permit for a demolition.
- `detroit-311.csv` : Each record represents a 311 call, typically a complaint
- `detroit-crime.csv` : Each record represents a criminial incident.

For each incident type, the location is included as a latitude, longtiude pair and various timestamps are included, typically in the format mm/dd/yyyy hh:mm:ss PM, for example 03/11/2015 04:23:11 PM.

Understand the Domain

Typically you will want to interact closely with domain expert stakeholders, but in some cases you may be asked to brush up on a new domain on your own.

Find some articles on blight and abandonment, the problems they cause, actions typically taken by city planners, prior approaches to prediction, and any equity or ethics considerations.

Some articles you may find useful:

- Spatial distribution of abandonment
- Early statistical approaches to predicting blight
- The relationship between abandonment and crime
- Detroit demolishes its ruins: 'The capitalists will take care of the rest'

Discussion Prompt: Share your background, interest, and goals for this Capstone Project, and any questions or considerations from your domain research. How important is this problem? How accurate do you think the models will be? What kinds of concerns might there be around equity? For example, in some cities, 311 calls may be rare in poor neighborhoods, so a model that predicts abandonment that uses 311 calls may favor certain neighborhoods over others.

Week 2. Create a list of "buildings" from a list of geo-located incidents

All data files consist of incident data: each record is an incident which occurred at a particular time and place.

The location of each incident is provided as geo-coordinates: a (latitude, longitude) pair. The project involves classifying the blight risk of specific buildings, so a first step is to parse the geo coordinate data and use these coordinates to try to cluster all incidents that occurred, ideally, at the same address. Note that the data is messy! You'll need to be creative to try to assemble your best guess of what constitutes a building.

Plan to spend some time visualizing the incident data to gain an intuition about the problem. In particular, consider creating maps using, say, Google maps for displaying a specific lat/lon location or CartoDB. You may of course also want to create non-spatial visualizations using, say, Google Fusion Tables or Tableau, or libraries such as vega-lite, bokeh, matplotlib, or R. Posting your visualizations online and discussing them in the forum is encouraged!!!

Milestone:

1. Create a file where each record represents a building. In most approaches, each building will be associated with some spatial extent so that you can determine which incidents will be assigned to it. One way to define a spatial extent to encode a rectangle centered on the building, such that your records are of the form building_id, lower_left_latitude, lower_left_longitude, upper_right_latitude, upper_right_longitude. You might also use a circle by associating each building with a center and a radius. Whatever you choose, you may need to consider how to handle incidents that fall outside of any bulding extent or incidents that fall within multiple building extents.

2. Write a function that, given a latitude and longitude, returns one or more buildings associated with that location. This function will use the information in the building file you created.

Assignment: Answer the following questions and submit for peer review. Your peer rviewers will offer suggestions and feedback.

1. What tools are you using? Are you using any special libraries to work with spatial data? How are you visualizing the data?

2. What outliers are you noticing? Consider what might be causing these outliers, and how you might specially address them.

3. How did you use the geocoordinates to define a "building"? How might you need to change your approach if you were working with 100x the number of incidents?

4. How might your solution change if conditions of the problem change? Would you face scalability issues if you were working with 100x the number of incidents? Would high-density areas like Manhattan in New York City make differentiation between buildings more difficult? Are there heuristics used about how buildings tend to be arranged geometrically that may be violated in certain places? None of these issues are necessarily a problem, but the more general the solution, the more likely it is to be used in additional cities.

5. Did you treat all incident types the same in determining whether to include them in the temporal history of a building? If not, why not?

Milestone 3: Construct a training dataset

To train a supervised model, you need ground truth labels. That is, you need a set of buildings that are pre-labeled as either blighted or not blighted. And you need about as many positive examples as you do negative examples, or your model will not perform well. There will be many, many more non-blighted buildings than blighted, so you will need to randomly select non-blighted buildings to include, and your model might be sensitive to this sampling.

For Milestone 2, you used the set of incidents across all files to create a list of buildings.
Each record in the file detroit-demolition-permits.tsv represents a permit issued in the city of Detroit. You may assume a building is blighted if at least one permit incident marked 'Dismantle' can be assigned to that building.

You will be ignoring all 'Dismantle' incidents when evaluating your model. (Why?)

Milestone:

1. Produce a file where each record is of the form building_id, label, where label is either blighted or not_blighted. Ensure that about 50% of the buildings are blighted and 50% are not. We have not created the features yet, so this file is not useful, but this step will ensure you can extract the labels, and you will follow a similar pattern for the features.

Assignment: Answer the following questions and submit for peer review.

1. Why is it in important to ignore the 'Dismantle' incidents during evaluation?

2. Could our labels be incorrect? If you were paid to pursue this project, how might you double check that your labels were accurately reflecting ground truth?

3. How many blighted buildings did you come up with?

Week 4: Train and evaluate a simple model

You will build a tirival model based on a single feature: the number of records in `detroit-blight-violations.csv` associated with that building.

For each record in `detroit-blight-violations.csv`, find the corresponding building and increment a counter.

Milestone:

Using scikit-learn in python (or R, if you wish), write a program to train a model and evaluate it using 5-fold cross validation. Evaluate your model using its accuracy. Use regression, a decision tree, or any other appropriate model.

Assignment: Answer the following questions an submit for peer review:

1. What method did you use to learn the relationship?

2. Does this model seem to work? How well?

3. What other properties of a building might help predict blight?

4. Is accuracy the only way to evaluate the model?

Week 5: Feature Engineering

For each building, you need to extract a richer set of features from the incident data and construct a feature vector. The feature vector will include at least two columns: the building_id and the label. You will create many additional columns that the model will use to predict the label. These features may be numeric, text-derived, or otherwise computed. Feature engineering is what makes or breaks analytics projects, not the method.

Your results will be highly dependent on the features you select -- you should expect to return to this step as you iterate.

Assignment: Answer the following questions and submit for peer review.

1. What features did you use? How did you select them? Did you leverage a corpus wide normalization? If so, why?
2. What types of features would you pursue if you had more time? Why do you think these features might improve the accuracy of your model?

Week 6: Extensions

Using data.detroitmi.gov or other inline sources, identify additional datasets from which you can derive additional features that may help you improve your model.

Assignment: Submit a brief report describing your features, your method, and the accuracy of the model. If you believe your solution works especially well and would like it to be reviewed by stakeholders in the City of Detroit and Socrata, upload your code to a public github repository and provide the url here.