LA City Health Prediction

Yifeng Xie

University of Southern California, Los Angeles, CA 90007

Abstract

In this paper, data of abandoned waste collection activity from LASAN is used to predict LA city health status. I hypothesize that waste collection duration (between the start time and the completion time of a collection request) is the most important variable in the dataset towards health status, which is proved true with the help of t-test analysis and decision tree model.

Introduction

Nowadays, people living in urban areas care more about health, which is related to many other aspects like sanitation, household income and even crime records. I use waste collection data to predict LA city health status. Mapbox is applied to implement waste collections' map visualization. NodeJS is used to preprocess data, aggregation data and generate geojson files for map visualization. After t-test analysis on the data, I assume that waste collection duration is the most important variable towards health status. To prove it as a true statement, a machine learning method named decision tree is employed. From the t-test results and the decision tree model, it is easy to find that average waste collection duration and quantity of electronic waste are two variables contributing more to the health status.

Raw Dataset Description

The text below a second-level heading begins without indentation. Use of the subsection heading style is required.

The raw dataset used for machine learning and predicting LA city health status is named abandoned waste collection activity, which is from LASAN. There are total 16614 instances, which take places within the whole November of 2015. This dataset has 8 attributes -- service request number, service request type, service request status, creation date, scheduled date, yard, council district and completion date.

After deleting some useless attributes like service request number and yard, service request type, creation date, council district and completion date are four variables that I used for the next step.

Since it is not easy to rate health status directly, I choose median household income data from Los Angeles City Council Districts Economic Report to represent health status. To make things simple, three levels of health status are set up. Based on Fig. 1 below, we can assume that within LA city, District 11, 12, 3, 5, 4 are the healthiest districts, which rank Level 1, while District 2, 7, 6, 15 rank Level 2, and District 13, 1, 14, 10, 8, 9 rank Level 3.

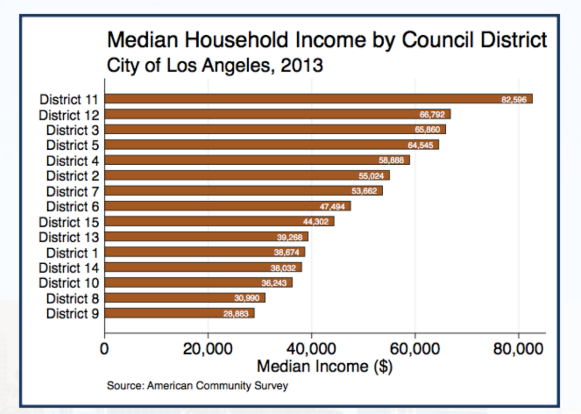


Fig. 1 Median household Income by Council District in LA city in 2013

Data Preprocessing

I set service request type and duration of a request as variables to generate a clean new dataset. In the dataset, there are totally six attributes – bulky items, electronic waste, illegal dumping pickup, metal/household appliances, duration and health status. The former four are from the service request type. For example, if a request is about bulky items, the instance has a value of 1 for bulky items and a value of 0 for the other three attributes. The duration attribute is generated from the creation time and completion time. Because the raw dataset is within only a month, some instances lack completion time. I make up the missing data and follow the rules below. If the creation time is after

November 26 and the completion time is missing, I make up the completion time by assuming the duration is 5 days, since the average duration is 5 days. If the creation time is before November 26 and the completion time is missing, I make up the completion time just as December 1.

To predict each council district's health status, I aggregate the data above by council district. Quantities of bulky items, electronic waste, illegal dumping pickup and metal/household appliances of each council district are aggregated as four new variables. All duration is another variable by adding up all the duration of each district. I set up two more dependent variables named all collections and average duration, which are quantity of all types of waste and average duration of each council district respectively.

Map Visualization

To plot the information from the data to a map, I choose Mapbox as my data visualization tool. It is a useful tool that is implemented with HTML, CSS and JavaScript code. It also needs geojson files to plot geolocation information. To fulfill this request, I use NodeJS to preprocessing the raw data and generate some geojson files which contain geolocation information of all these waste collections.

The visualization results are as Fig. 2 and Fig. 3.

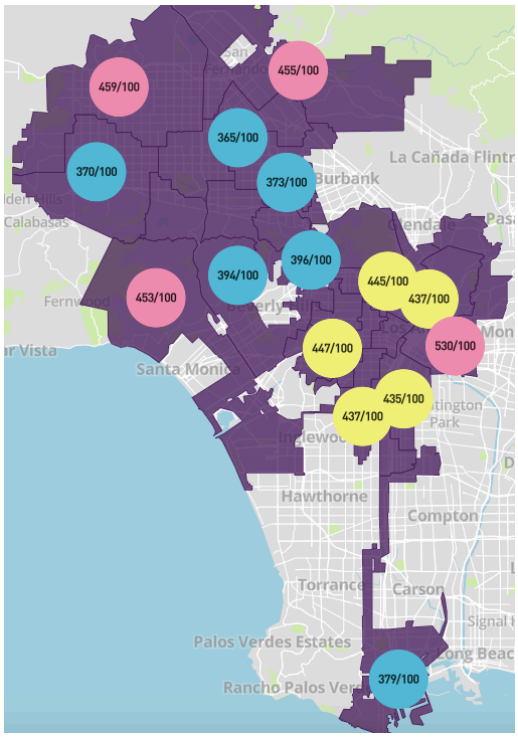
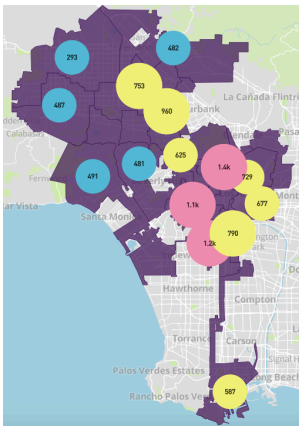


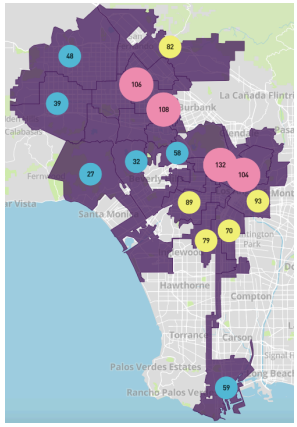
Fig. 2 Map visualization of average duration by council district

My hypothesis and T-test

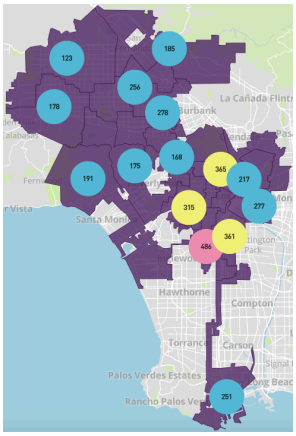
To figure out the relationship between collection types and duration with the health status, I hypothesize that waste collection is the most important variable towards health status and employ t-test to prove their correlation. The t-test results are as Tab. 1.

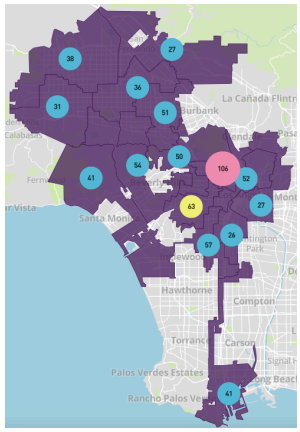


(1) Bulky items



(2) Electronic waste





(3) Illegal dumping pickup (4) Metal/household appliances

Fig. 3 Map visualization of four types of waste collection by council district

P-value	Bulky Items	Electronic Waste	Illegal Dumping Pickup	Metal/Household Appliances	duration
Health Area 1 vs Health Area 2	0.863897139	2.05E-06	0.777530701	3.28E-06	0.0023781
Health Area 2 vs Health Area 3	0.172292712	0.000268404	0.571759113	0.642751576	8.37E-22
Health Area 1 vs Health Area 3	0.273174323	0.055348648	0.832505132	2.14E-06	3.47E-08

Tab. 1 P-value of T-test

From the table above, we can conclude that electronic waste and duration are two major variables towards health status.

Machine Learning -- Decision Tree

To prove my hypothesis and predict health status, a machine learning method – decision tree is implemented towards the data. I make use of Jupyter Notebook as my machine learning IDE and Scikit Learn as my machine learning library. To avoid deviation, cross validation is applied.

For the non-aggregated data, I finally get a prediction accuracy rate of 0.52, and a decision tree model as Fig. 4.

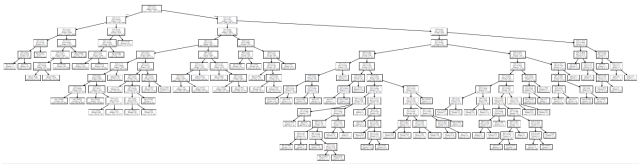


Fig. 4 Decision tree model for non-aggregated data

The top one is as Fig. 5. In Fig. 5, X[4] is the duration.

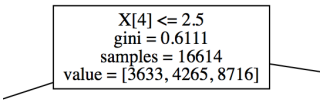


Fig. 5 The very top label of Fig. 4

For the aggregated data, I finally get a prediction accuracy rate of 0.63, and a decision tree model as Fig. 6. In Fig. 6, X[1] is the quantity of electronic waste, X[5] is the all duration per council district.

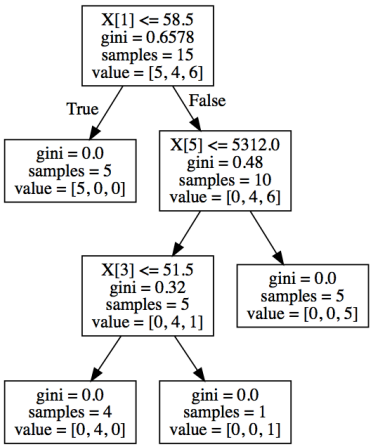


Fig. 6 Decision tree model for aggregated data

From Fig. 5 and Fig. 6, we could draw a conclusion that quantity of electronic waste and all duration per council district are two most important variables to predict health status. The less collections of electronic waste the district has, the healthier the district is. The less total duration the district has, the healthier the district is.

Conclusion

In the paper, I firstly find a dataset about waste collection to predict LA city health status. Then I preprocess the raw data and aggregate it by council district, as well as implementing map visualization. As the next step, I make a hypothesis that duration is the most important feature that contribute most to predict health status. To prove that, I finally make a t-test to get its p-values, and apply decision tree algorithm to obtain decision tree models.

Based on the p-values and decision tree models, I successfully prove my hypothesis as a true statement and draw a conclusion that both quantity of electronic waste and all duration of each district are two most important features and they have a negative correlation with health status.

References

https://data.lacity.org/A-Livable-and-Sustainable-City/LASAN-Solid-Resources-Abandoned-Waste-Collection-A/97ra-aqza/data

http://www.lachamber.com/clientuploads/policy_issues/15_BeaconReport_Web.pdf

http://scikit-learn.org/stable/modules/tree.html

https://www.mapbox.com/mapbox-gl-js/example/cluster/