Analysis of 311 hotline resolution

Factors that influence the number and type of cases and their resolution status

Bozhi Hao, Hongzhou Tao, Kaiyang Li, Yixun Tian
Data Science and Analytics Program, Georgetown University

Introduction

311 service is a non-emergency city service hotline. Although the issues it deals with are not urgent, these everyday cases, including street light maintenance, garbage collection and so on, are crucial to citizen's quality of life.

To better serve the public, municipal governments are committed to more rationally allocating resources and solving problems. For this purpose, our analysis of the 311 dataset aims to figure out the factors that influence the number and types of issues, and how well they are resolved. This study focuses on the field of municipal service and has guiding significance for public service policy and residents' lives.

Analysis and methods – About the Data

This 311 dataset contains the records of 311 city service hotline of san francisco. It has 519919 observations (rows) and 27 columns (26 attributes), including service request information and other basic information about different neighborhoods in san francisco in 2012, 2014, 2016 and 2018. The attributes include date, case, average income, population, race proportion, resolution time, location, turnout time and so on. Below are some features that needs explanation:

mobile.dumm	1/0: Request using/not using mobile phone
photo.dumm	1/0: Case description with/without photo
white/black/asian	Proportion of certain race in the neighborhood in certain year
turnout_one/two_lag	The turnout time of the service request one/two before the current one

Before we analyze the data, we did some data processing and preparation:

Filtering NA values → Converting to dummy variables → Selecting features and rows → Merging or disassembling features → Calculating needed features

Analysis and methods – Exploratory Data Analysis

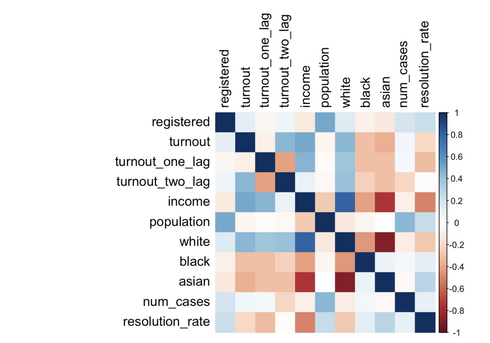

Fig-1: Correlation matrix of some numerical features of different neighborhood in San Francisco

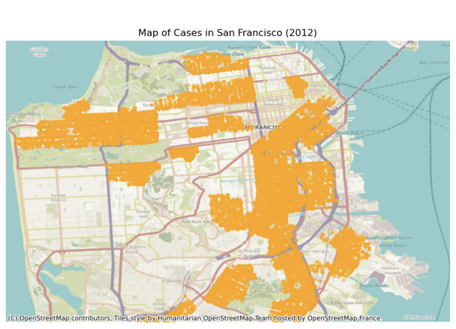

Fig.2 - Geographical distribution of cases in 2012 in San Francisco.

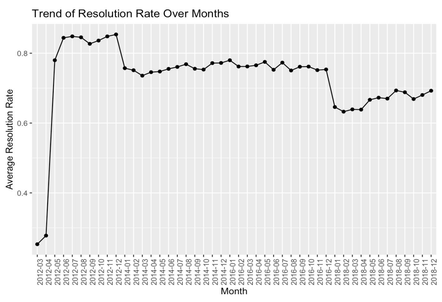

Fig-3: Line Graph of the trend of resolution rate over time, with month as unit.

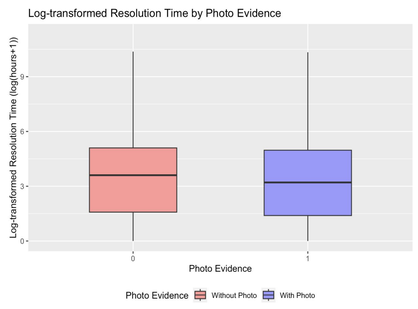

Fig-4: Boxplot of Log-transformed Resolution Time (log(hours+1)) by Photo Evidence

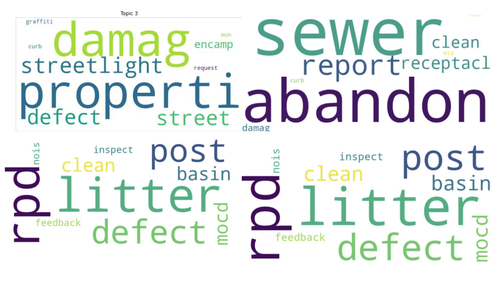

Fig-6: Word Clouds of different neighborhoods (in acronym)

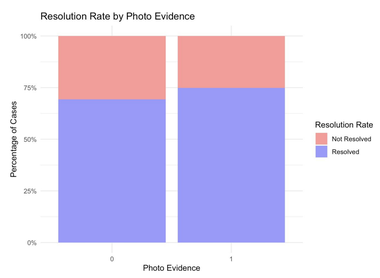

Fig-5: Stacked columns of Resolution Rate by Photo Evidence, with y axis as percentage

Analysis and methods – statistical methods and models

1. Hypothesis test

Hypothesis test is a method to validate whether there is enough evidence to reject the null hypothesis (often like"A has no influence on B" or "A is the same as B"). In this research, hypothesis testing is used to figure out:

1.1 Whether racial composition is influencing resolution time. This is done by fitting a linear model in which X is the proportion of each races and y is the resolution time, then calculated T-statistic = coefficient/standard error, and p-value = 2*(1-CDF(t)).

1.2 Whether cases with photo evidence result in shorter resolution time. This is done by using wilcoxon test to compare the medium.

1.3 Whether cases with photo are more likely to be resolved. This is done by using Chi-squared test to analyze whether there is statistical independence between the two variables.

2. Negative Binomial Model

Negative Binomial Model can analyze overdispersion data, it is used to analyze the log change in the response variable (num_case) with each category (neighborhood, case_type), compared to the baseline category (bayview, 311 external request). The larger the Z absolute value, the more evidence against null hypothesis.

3. Subset selection

3.1 What are the factors that influence the number of cases.

3.2 What are the factors that influence the resolution rate.

Both questions requires subset selection.

In this section different subset selection methods are used[1]:

- Best subset selection: An exhaustive method that attempts all possible combinations of variables to select the best performing one.
- Forward stepwise selection: Begins with an empty model and adds variables incrementally.
- Backward stepwise selection: Begins with a model that contains all variables and removes variables incrementally.
- K-fold Cross Validation: A method used to evaluate model's performance. Helps to prevent overfitting.

4. Regression

In this section we tried linear model, general linear model, ridge regression and lasso regression.

Results

1. Hypothesis test

1.1Whether race proportion is influencing resolution time.

	coefficients	standard errors	t-statistic	p-value
White	3302.559826	231.615099	14.258828	1.337552e-11
Black	5596.298518	1492.118149	3.750573	1.354206e-03
Asian	2878.458237	369.738635	7.785116	2.505272e-07

All null hypotheses are rejected.

1.2 Whether cases with photo evidence result in shorter resolution time.

w = 3.5024e=10, p-value<2.2e-16, the null hypothesis is rejected.

1.3 Whether cases with photo are more likely to be resolved.

X-squared=1925.1, p-value<2.2e-16, the null hypothesis is rejected.

2. Negative Binomial Model

	Estimate	Std.error	Z value
neighborhood: Mission	-1.25747	0.17085	-7.360
case: MOH	-5.10384	0.37518	-13.604

There are neighborhoods and types of cases that result in quite different num_cases. And there are strong evidence against null hypothesis.

3. Subset selection

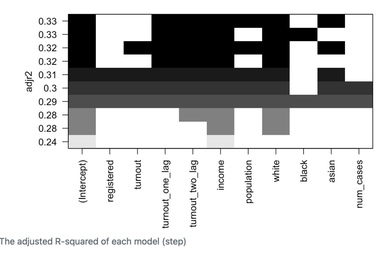

Fig-7: The adjusted R-squared of each model (step)

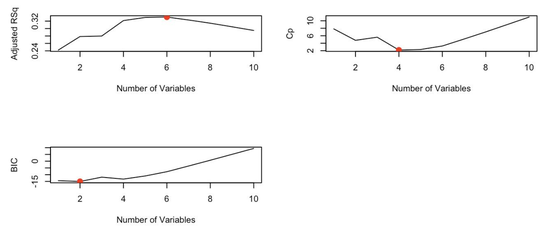

Fig-8: Best subset according to Adjusted R-squared, Cp, BIC respectively.

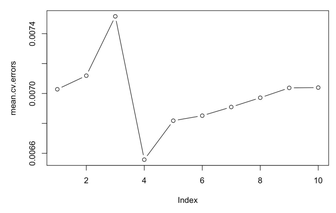

Fig-9: Best subset according K-fold cross validation

problem	Adjusted R-squared	Cp	BIC	k-fold CV	selected
3.1	7 features	6 features	1 feature	3 features	7 features
3.2	6 features	4 features	2 features	4 features	4 features

The figures shown are for question 3.2, 3.1's figures are just alike.

4. Regression

MSE of different regression methods: (glm() is the same as lm(), indicating the data just follows linear pattern)

	Linear model	general linear model	ridge regression	lasso regression	selected
3.1	73144749	73144749	86177413	86394589	Linear
3.2	0.004052783	0.004052783	0.003238037	0.004021761	Ridge

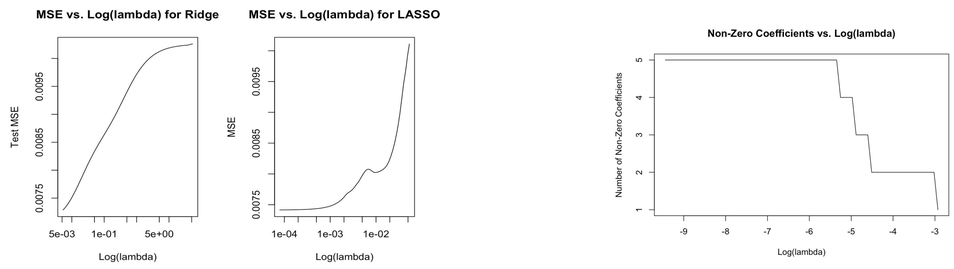

Fig-10: Parameter selection for ridge regression and lasso regression. Test MSE vs log(lambda) for ridge and lasso regression.

Fig-11: Number of Non-Zero Coefficients vs log (lambda) for lasso regression.

Conclusion

- The main factors influencing number of cases: income, population, race proportion of white, black and asian, turnout time(may indicate the resource richness). Different neighborhoods are likely to have different types of cases.
- The main factors influencing resolution rate: income, turnout time, race proportion of white.
- Racial composition do influence case resolution time. And resolution time vary with moth and year.
- Having a photo in description helps with resolving the problem faster and more successfully.

Reference

[1] code source: https://jfh.georgetown.domains/dsan5300/