Aparna Shastry

Irvine, CA

Predicting Building Permit Issuance Times for San Francisco!

February 23, 2018

Overview

This document proposes a data science project that uses building permit related data obtained from open data portals of the city of San Francisco. We study the data and draw a few interesting inferences.We aim at modeling the time taken to issue building permits by Permit Services wing of Department of Building Inspection (henceforth called DBI), as accurately as possible to predict the expected wait times for newly filed permit applications

Goals

1. **Learning the Process:** Starting from thinking of interesting ideas for a possible client, all the way to presenting the work (to even non-technical person), learning the data science process is my first and foremost goal.

2. **Use the data to come up with answers :** The data can provide answers to the following questions:

a) How to interpret the records with zero or very small values for “Estimated Cost” or “Revised Cost”? Is the “Revised Cost” always more than “Estimated Cost?”

b) What is the best day of the week to visit DBI, to file an application form? Is the popular belief “mid-week (Wednesday) is the least crowded and hence best to visit government or city agencies” true in this case?

c) What type of permits are issued on the same day? Which types take least time?
d) Is there any particular quarter of each year which has higher application counts or average wait times? Can it be justified from the business knowledge?
e) Is there a statistically significant difference between issuance times for Residential Vs Commercial buildings?
f) Is there a statistically significant difference between wait times of fire only permits and not fire only permits? Similarly with site permit. What is the interpretation?
g) What is the trend across the years with respect to number of applications or wait times? Is there any anomaly?
h) What are the main factors influencing the building permit issuance times?
i) How does the building estimated cost relate to wait times?
j) Is there a correlation between the location and wait times?

We are aware that the data may not provide the conclusive statements to all of these, and it might give us some surprising insights as well. We are curious to learn more and more from the data about the nature of issue times.

3. **Make a model to predict the expected delay in getting building permit**: Try fitting different Machine learning models, tune the hyperparameters and see which one closely mimics the permit issuance process. Evaluate predictive power/accuracy of the model.

Possible Clients

1. Builders/Planners
2. Agencies that apply for permits on behalf of owners/builders
3. Real estate companies

Approach

1. Download data from the portals for 2013 to present for San Francisco city

2. Clean up to fill the nulls, convert to datetimes etc
3. Do EDA and make initial observations
4. Conduct hypothesis tests
5. Check if the problem still falls into “those that need predictive models” category. If yes, or if no, document it and justify
6. Whatever may be the result of 5, consciously decide to go ahead to fit machine learning models . If the answer to 5 was yes, really great! Otherwise at least the modeling process would be a learning experience.
7. Document all the steps and make final conclusions/recommendations

Data Source

1. Building permit data from San Francisco City : This data is updated on weekly basis. It has 43 columns and over 1 Million rows starting from the year 1968. Many columns are not very useful. Excluding location info, we could extract 16 predictors that would help in answering the above questions.

Milestones

1. Capstone Milestone Report Submission:

   Submit the Milestone Report by March 5th, 2018

2. Final Submission

   Submit code, report and power-point presentation by March 15th, 2018

Deliverables

1. IPython Notebook
2. Power-Point Presentation
3. Report
4. Blogpost