House Price Prediction in Natural Hazard Prone Areas

Capstone Project 1: Proposal

1. What is the problem you want to solve?

Goal:

Given a property details and natural hazard attributes, predict the median house price in San Jose city using Zillow property data and natural hazards data such as earthquake, landslide and fire hazards. The main goal for this project is to find major attributes for the contribution of house prices in San Jose city.

Background:

California is well known for Earthquakes and it is overdue for a major shaking. In 1989's Loma Prieta earthquake, the world saw what liquefaction can do, as blocks of the Marina District were reduced to rubble. Some properties are on top of fault lines like the ones damaged in South Napa that had fault traces running through them. When the ground actually splits during an earthquake, it can damage buildings and utility lines far beyond what shaking can do. Properties that are prone to "ground failure" are required by law to disclose that information to potential buyers which could potentially change buyer’s decision scenario.

Apart from earthquake hazard, 2018 wildfires have ravaged more than 1.2 million acres, destroyed more than 1200 homes in California. Housing market changed after that devastation in Northern California.

Properties that are prone to "ground failure" are required by law to disclose that information to potential buyers -- but the information isn't always easy to find until the transaction is already underway.

2. Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn’t have done otherwise?

Clients:

This study has been motivated by the need for a variety of stakeholders, including residential valuers, real estate agents, policy makers, insurance providers and mortgage lenders, to be able to accurately estimate the way in which changing perceptions of the risk of damage could alter property values.

Benefits to Clients:

(a) Real estate websites:

For example, if this study is made available to real estate companies such as Zillow or Trulia, then it would result in following business value additions:

- Zillow customers would be more informed on natural hazard attributes before buying a property.
- Zillow customers would use this hazard attributes as a key element, while making a decision on property purchase.
- Estimating the value of property by including various hazard attributes would yield higher precision of forecasting, which in turn leads to customer satisfaction and retention for the firm.
- Number of customers visiting Zillow would increase due to inclusion of hazard attributes compared to its competitors.
- Zillow makes money through Ad Sales as well. So providing natural hazard information for a property would invite advertisers from related companies such as seismic retrofit construction firms, earthquake insurance providers, etc to display their ads on the same page of those properties.
- Due to increase the volume of customers visiting the website, that result in increased number of premium real estate agents sign up for subscription services.
- With more customer engagement, cost per click from mortgage lender would get increased.

(b) Insurance Providers:

- With this information on natural hazard attribute tied up with property, insurance providers are less likely to encounter a loss on property claimed on those attributes. Because the insurance premium of the property would include this risk factor as well.

3. What data are you using? How will you acquire the data?

House Data:

For this project, Zillow property data for San Jose city will be collected and the method is outlined below:

- Zillow property data:
  Planning to use web scraping to get access the Zillow property data using python and web scraping packages such as selenium and BeautifulSoup.
  https://www.zillow.com/homes/san-jose_rb/

Natural Hazards Data:

In this project, natural hazards data such fault zone, landslide, liquefaction and fire hazard data will be collected from below sources:

- Earthquake (fault, liquefaction and landslide zones) hazard data: https://spatialservices.conservation.ca.gov/arcgis/rest/services/CGS_Earthquake_Hazard_Zones/SHP_ZoneInfo/MapServer
- Wildfire hazard potential data: http://www.fire.ca.gov/fire_prevention/fhsz_maps/FHSZ/santa_clara/San_Jose.pdf

Data Collection:

House properties data from Zillow:

- Zillow website API has inbuilt limitation on extracting data from Zillow. Therefore, web scraping will be done to extract the fields for single family homes and townhouse for “sold” and “for sale”(including By agent, By owner, New construction, Foreclosures, Coming Soon).
- Planning to extract property attributes such as latitude, longitude, address, zip, bedrooms, bathrooms, sqft, lot_size, year_built, price, sale_type, zestimate, date_sold, days_on_zillow, house_type and url using web scraping packages.

Earthquake Hazard data from CGS(California Geological Survey)

- CGS has recently developed web app for predicting earthquake hazard data for California. In this project, web app, sql, python requests and json packages will be used to extract earthquake hazard data from hazard web app.
- Zillow website API has inbuilt limitation on extracting data from Zillow. Therefore, web scraping will be done to extract the fields for single family homes and townhouse for “sold” and “for sale”(including By agent, By owner, New construction, Foreclosures, Coming Soon).

Fire Hazard Severity Zone data from California Fire Department

- Fire hazard data is available in pdf. San Jose city has very limited area comes under fire hazard severity zone. Parcel number and address will be extracted only for the area comes under fire hazard severity zone.

After collecting all property data and natural hazards data for San Jose, merge and append methods will be used to join data based on the address and parcel number attributes.

4. Briefly outline how you'll solve this problem.

Outline:

1. Data collection

   Tools: Python, selenium, BeautifulSoup, sql, python Requests, json packages, jupyter notebook

2. Data wrangling

   Tools: Python, jupyter notebook, Matplotlib, pandas, numpy

3. Exploratory analysis

   Tools: Python, jupyter notebook, Matplotlib, pandas, numpy

4. Visualization

   Tools: Python, matplotlib, folium, basemap and seaborn

5. Developing prediction models

   Tools: scikit

6. Summary and Report

   Tools: Jupyter notebook

Outcomes:

In this project, hazard and non-hazard properties in single family and townhouse will be analysed following segmentation approach.

- Geospatial visualization of natural hazard homes in single family and townhouse
- Most common features in single family (bedrooms, bathrooms,year_built, liquefaction, fault zone, landslide, fire hazard)
- Most common features in townhouse (bedrooms, bathrooms, year_built, liquefaction, fault zone, landslide, fire hazard)
- Distribution of number of houses with price bin for single family and townhouse
- Median number of bedrooms, bathrooms,sqft with price bin for single family and townhouse
- LIquefaction, fault zone,landslide, fire hazard with price bin in single family and townhouse
- Sold price distribution for various zip codes, hazards and non hazards for single family and townhouse
- Best month to put sale for single family and townhouse
- Most popular zip codes saleswise for single family and townhouse
- Number of homes sold from 2016-2019 for single family and townhouse
- Influence of natural hazard on sold price

- Median price/ sqft for hazard and non-hazard homes
- Median price trend over construction year
- Impact of liquefaction,landslides, fault zone, fire hazard on price of single family and townhouse
- People's comment on houses in hazard areas
- Correlation plot between features and price
- Influential features to predict house price

5. What are your deliverables?

Deliverables include code, detailed report and slide deck.