



Impact of Future Housing Developments in SF

Jasmine Teo, Sally Fang, Anton Bosneaga, Priya Reddy, Arsh Hothi

Table of contents

01

Motivation

02

The Data

03

Model Implementation

04

Conclusion

05

Biases & Future Work



Background & Motivation

ADUs

Housing in San Francisco

... is expensive.

AHDBP

Several factors have contributed to this problem:

- a limited supply of housing
- a thriving tech industry that has led to increased demand
- restrictive zoning laws that make it difficult to build new housing units.

Home
SF

SOMA

Research Question



Can we predict the value of new developments in San Francisco from historical property valuations?

San Francisco has implemented several policies and programs aimed at increasing the production of new housing units, particularly affordable units. These efforts have helped increase the supply of new housing in the city, but the demand for housing continues to outpace supply, leading to high prices and limited affordability.

Our research question for this project is whether we can predict the value of new developments in San Francisco based on historical property valuations.

The Data



Assessor Historical Secured Property Tax Rolls

DataSF – City of San Francisco's Open Data Platform

Includes all legally disclosable information (location, value of property, property characteristics) from July 1, 2007 to June 30, 2018 for San Francisco



Police Department Incident Reports

DataSF – City of San Francisco's Open Data Platform

This dataset includes incident reports that have been filed as of present starting 2013. These reports are filed by officers or self-reported by members of the public using SFPD's online reporting system.



San Francisco City Survey Data

DataSF – City of San Francisco's Open Data Platform

The City Survey asks residents to indicate their usage and satisfaction with city services and infrastructure like libraries, Muni, public safety, and street cleanliness. The City Survey was conducted every year from 1996 to 2004, and biennially from 2005 onward.

Data Cleaning: Housing Valuation

Encoded Zip Codes

Extracted and formatted coordinate data to join with San Francisco ZIP Codes.geojson using GeoPandas

Kept only the relevant features, and data from 2013–2019 (by the Closed Roll Year)

Selected Data

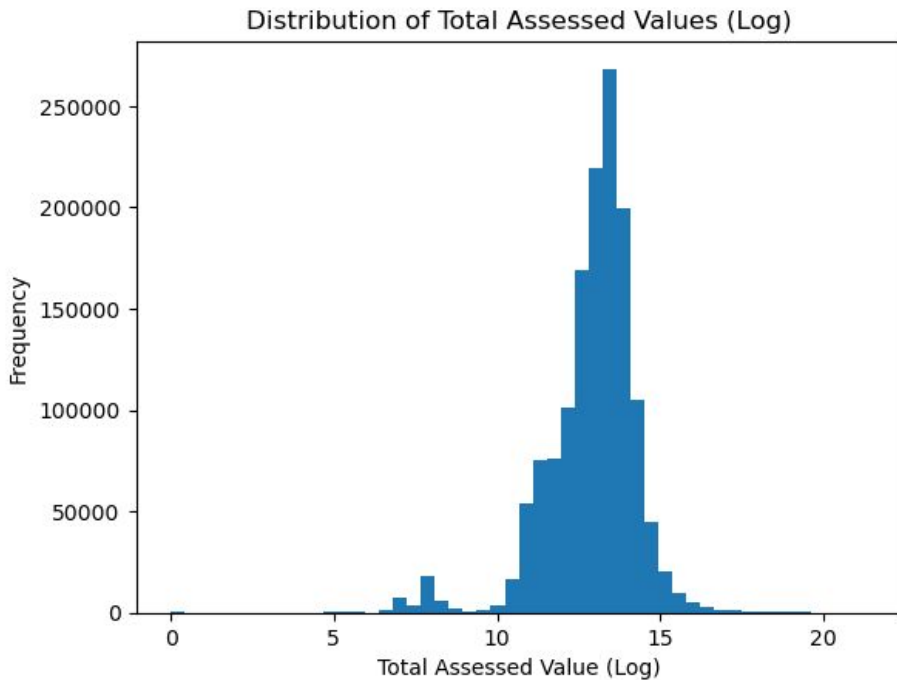
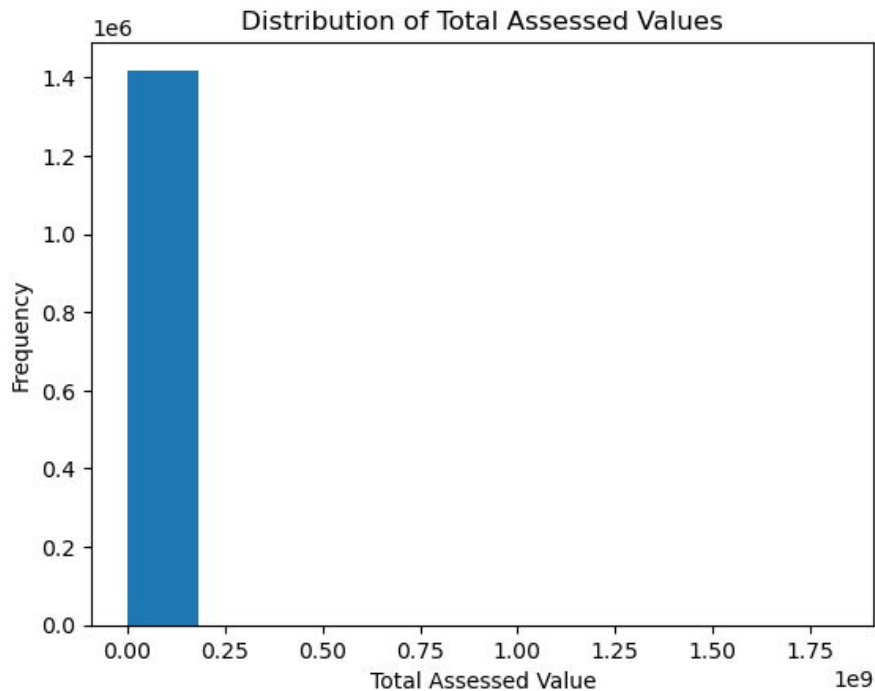
Calculated the Outcome Variable

Total Assessed Value
= Assessed Personal
Property Value + Assessed
Improvement Value

Valuation exhibits a right-skewed distribution because of a few expensive lots, we will instead look at the log

Logged Total Assessed Value

Total Assessed Values







Data Cleaning: Police Incidents



Combined 2 Datasets



Combined 2013-2018
and 2018-Present
together, dropped
duplicated data for
2018 using Incident ID
(unique identifier)



Kept only the relevant
features and columns
we are interested for
EDA and later
modeling

Selected Data

Encoded Zip Codes

Extracted and formatted
coordinate data to join
with San Francisco ZIP
Codes.geojson using
GeoPandas




Data Cleaning: City Survey



Selected Columns of Interest

Selected questions relating to cleanliness of the neighborhood, safety (both at day and night), household density, and zip code



Question about cleanliness changed over time, aggregated ordinal columns per response

Aggregated Features

Filtered by Valid Zip Code

Filtered for valid zip codes and years 2013–2019. Found that there are no zip codes available for 2015, and we only data of every other year

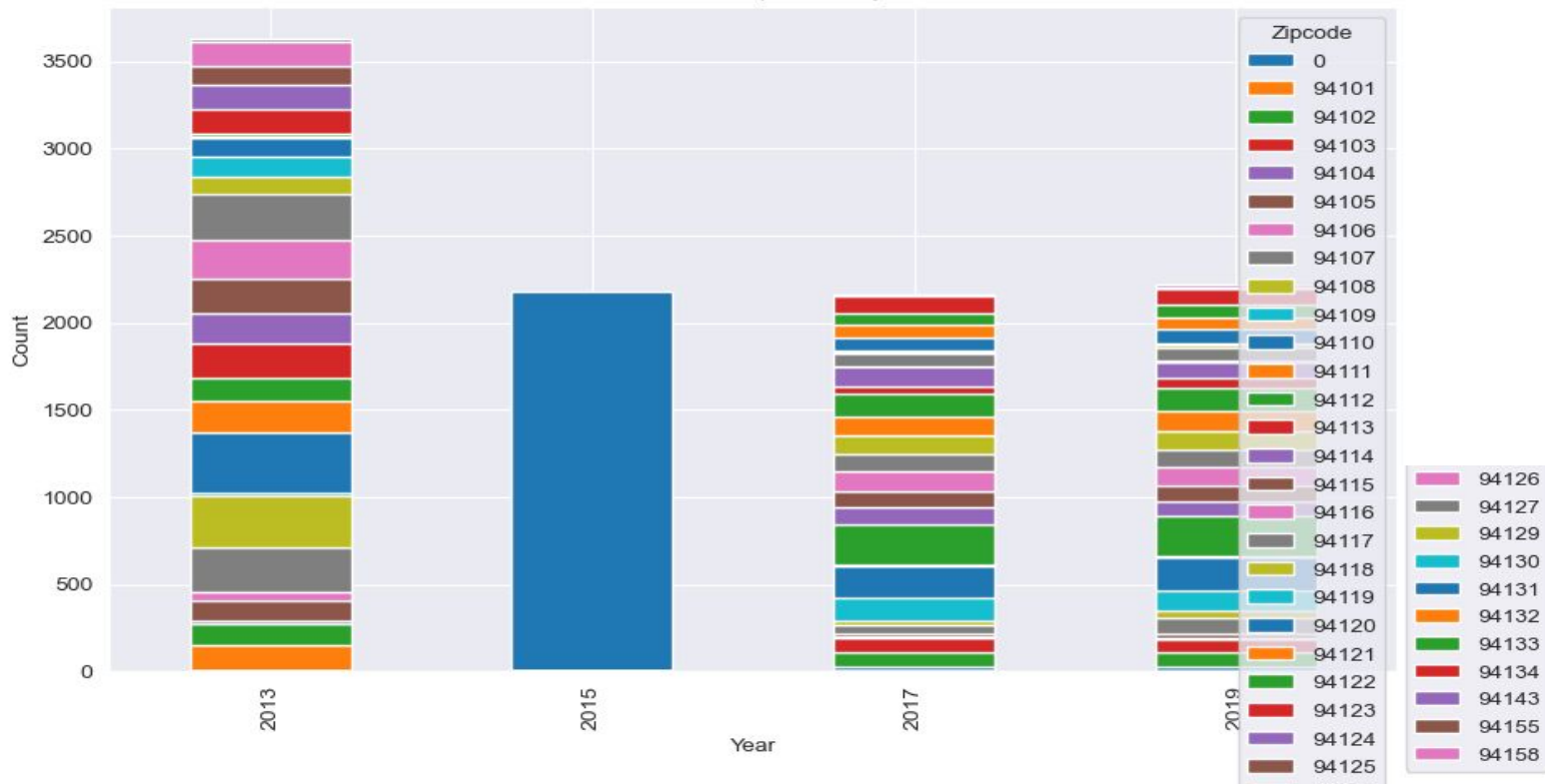
Filled in gap years with previous year data (i.e. fill in 2014 & 2015 & 2016 data with 2013 data)

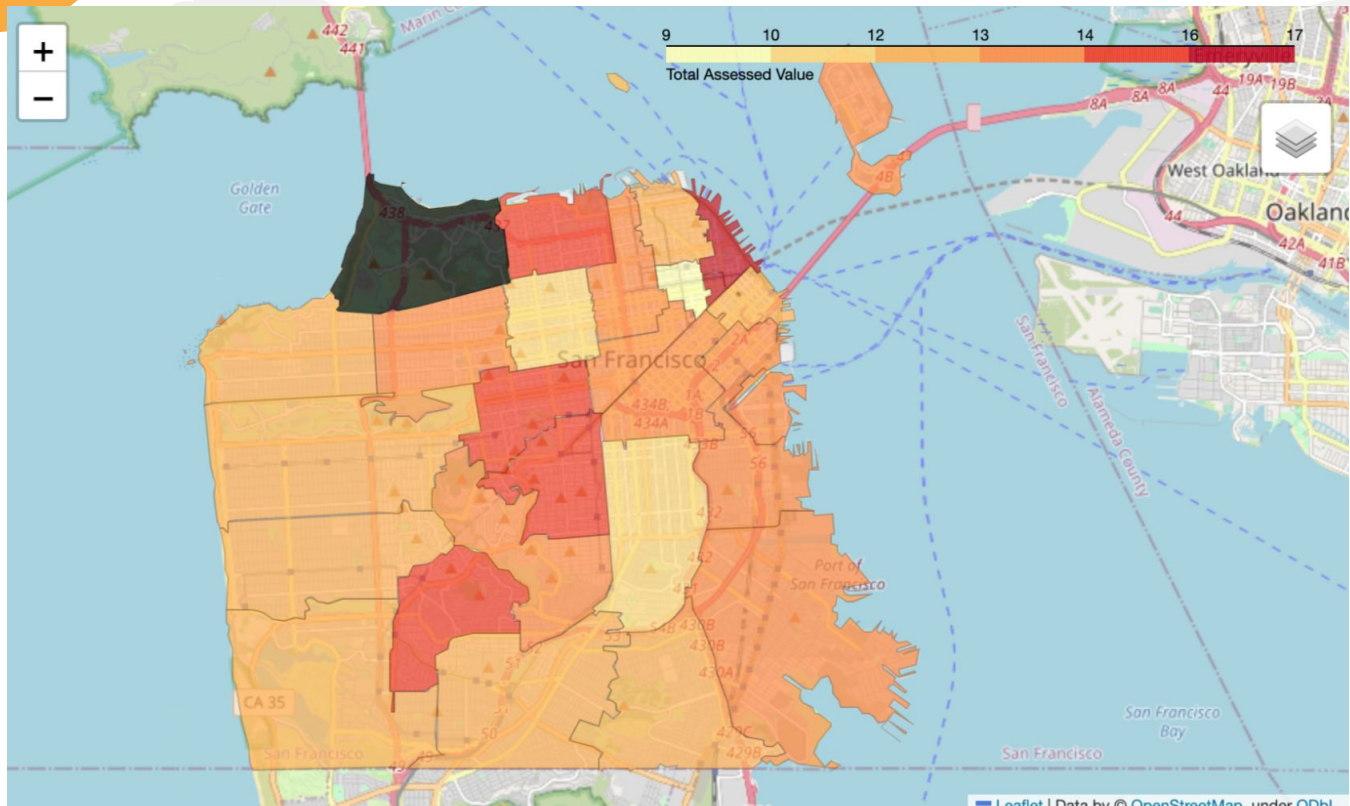
Filled in Gap Year Data

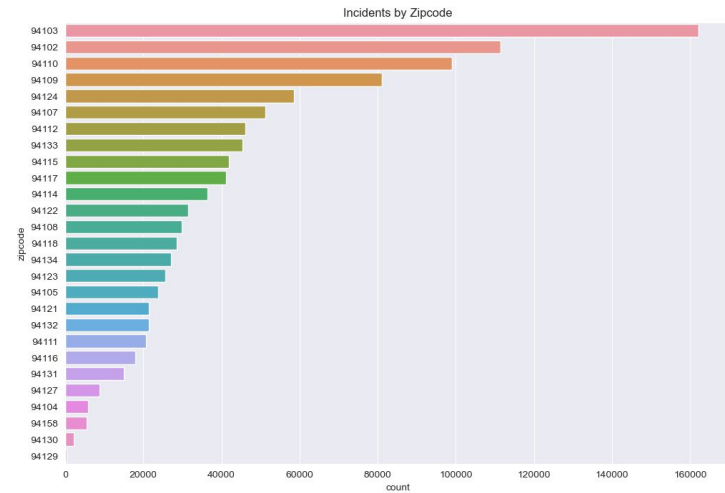
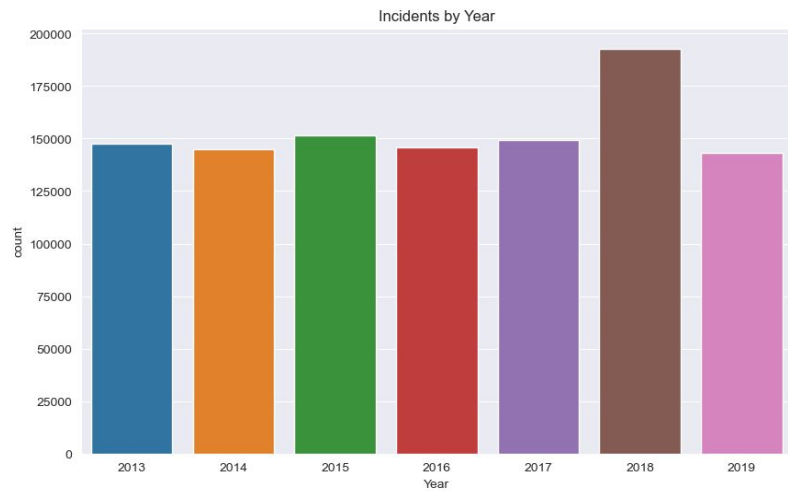


Distribution of Zipcodes

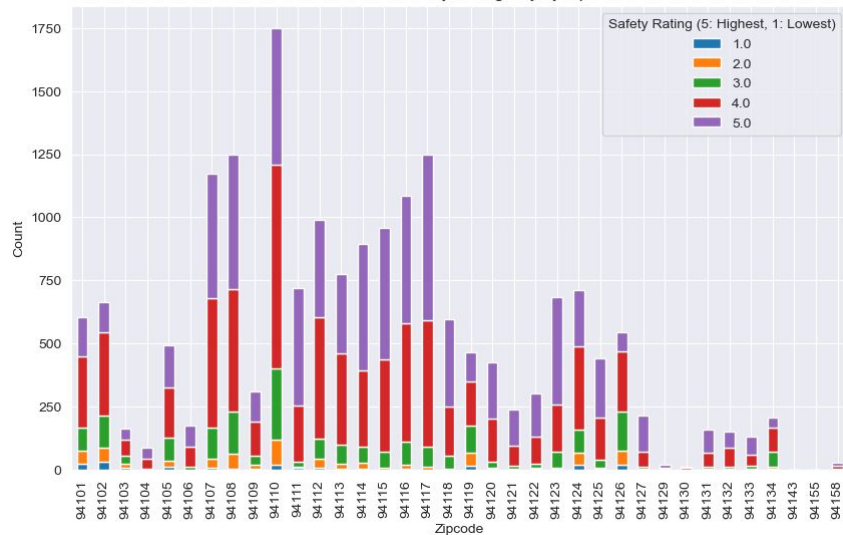
Distribution of Zipcodes by Year



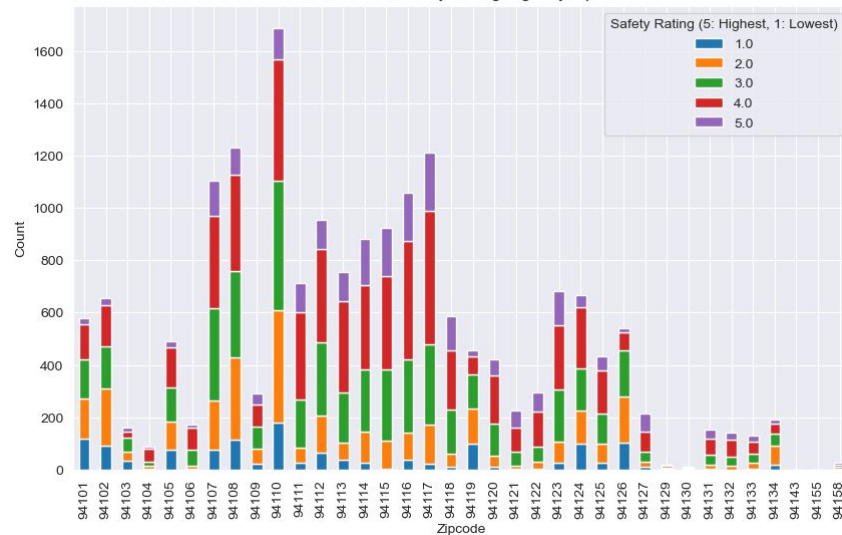




Distribution of Safety During Day by Zipcode



Distribution of Safety During Night by Zipcode



Putting it all together:

Merging the Data



Police Data

Year	Year of Interest
Zip code	Zip code crimes took place in
Incident Count	Number of Incident for that year and zip



House Value Data

Year
Zip Code
Bathrooms
Bedrooms
Stories
Units
+
+
+



Survey Data

Year	Year of interest
Zip Code	Zipcode of interest
Safe Day	Avg score for day safety
Safe night	Avg Score for night safety
Household Size	Avg Household Size
Cleanliness	Avg Score for cleanliness of Zip Code

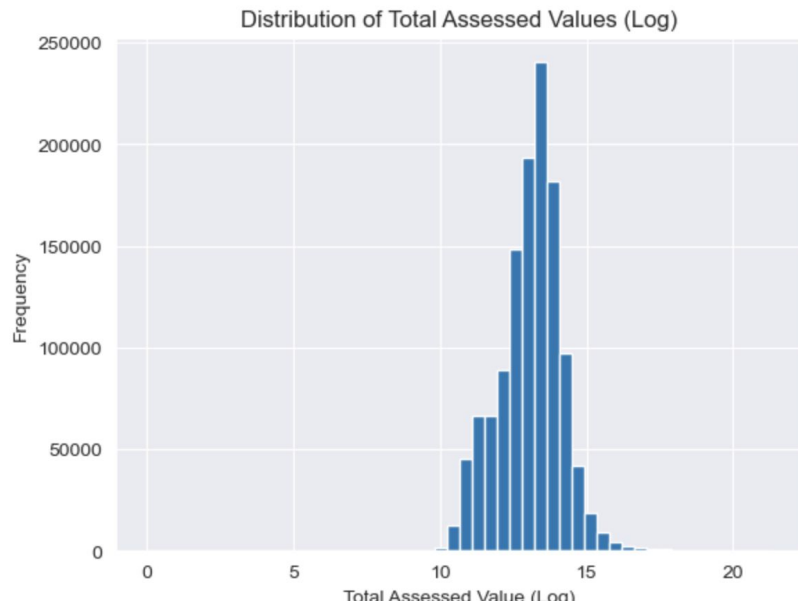
Multivariable Linear Regression

Why Linear Regression?

- Supervised learning model that has **high interpretability** used to predict continuous variables (such as price)
- Based on research we did on common methodologies for predicting housing prices it seems that linear regression models **are typically standard**
- Based on initial exploratory plots we can also see **linear relationship** between features and price

Implementation Steps

- Feature Engineering
- Split data into train, validation, and test sets
- Hyperparameter tuning
- Evaluate against baseline



Multivariable Linear Regression



Model Tuning

- Feature selection:
 - All Features
 - Only features with a correlation >0.5 or <-0.5
-
- Switch to Random Forest

Performance Metrics

Model	Train RMSE	Validation RMSE	Test RMSE
1. LM 1 (all vars)	339,933,381,452.719	308,289,496,854.8	NA
2. LM 2 (only high corr vars)	246,863,620,528.9301	254,280,097,924.99	252,689,331,870.50
3. Median Predictor (baseline)	7,850,186.20	9,624,777.43	7,221,440.69



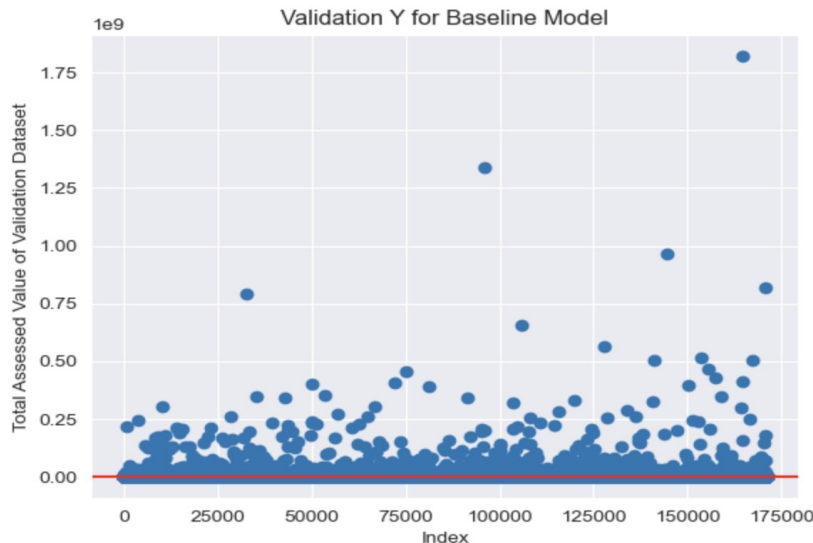
Baseline: Median Predictor

Poor Performance of Linear Model

- Use median of *Log of Total Assessed Value* as baseline predictor
- No hyperparameter tuning needed

Performance Metrics

RMSE on Train Data	7,850,186.20
RMSE on Validation Data	9,624,777.43
RMSE on Test Data	7,221,440.69



Random Forest Regressor

Why Random Forest?

- **Multiple decision tree models** built and each prediction is combined to make final predictions
- **Handles non-linear relationships** between features & target variable, *Log of Total Assessed Value*
- **Deals well with missing values and outliers**, providing robust predictions with relatively low risk of overfitting

Feature Selection

- Including irrelevant/redundant features can introduce **noise & unnecessary complexity**
- **Selected Features:** *Number of Bathrooms, Number of Rooms, Zipcode_94123, Number of Units, Zipcode_94105, Zipcode_94124, household_size, Zipcode_94112, Zipcode_94134, Zipcode_94122, Zipcode_94116, cleanliness, Year, Year Property Built, Zipcode_94104, Zipcode_94115, Number of Bedrooms, Zipcode_94109, Zipcode_94111, Number of Stories, Zipcode_94114, Zipcode_94103, Zipcode_94108, safe_day_binary*

Random Forest Regressor

Hyperparameter Tuning

- `n_estimators`: number of decision trees to include in the random forest
 - `n_estimators` = 10
 - `n_estimators` = 20
- `max_depth`: controls the maximum depth of each decision tree in the random forest
 - Creating deeper trees led to overfitting due to increased capture of complex patterns in the data → used default value instead

Performance Metrics

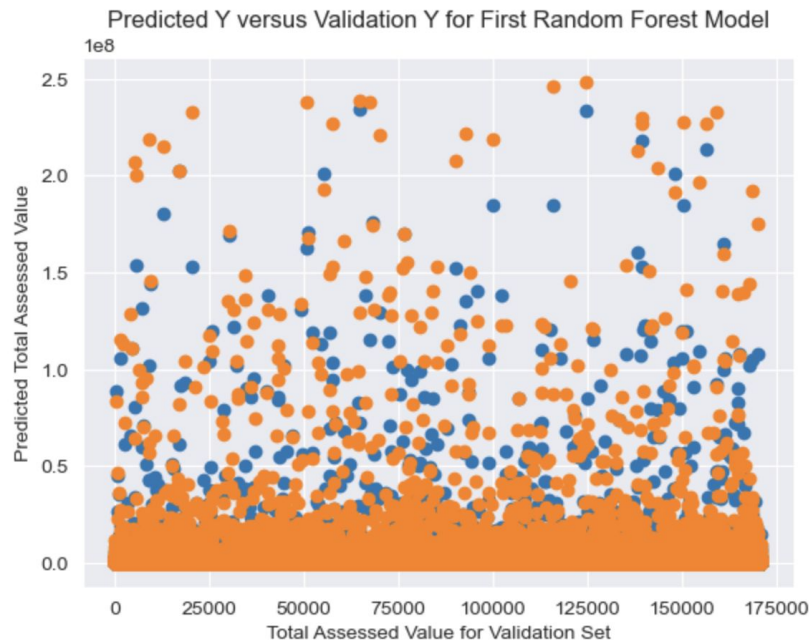
Model	Train RMSE	Validation RMSE	Test RMSE
1. RF (<code>n_estimators</code> = 10)	1,370,849.90	2,199,539.04	NA
2. RF (<code>n_estimators</code> = 20)	1,313,411.46	2,127,457.25	2,107,751.89
3. Median Predictor	7,850,186.20	9,624,777.43	7,221,440.69

← Final Model

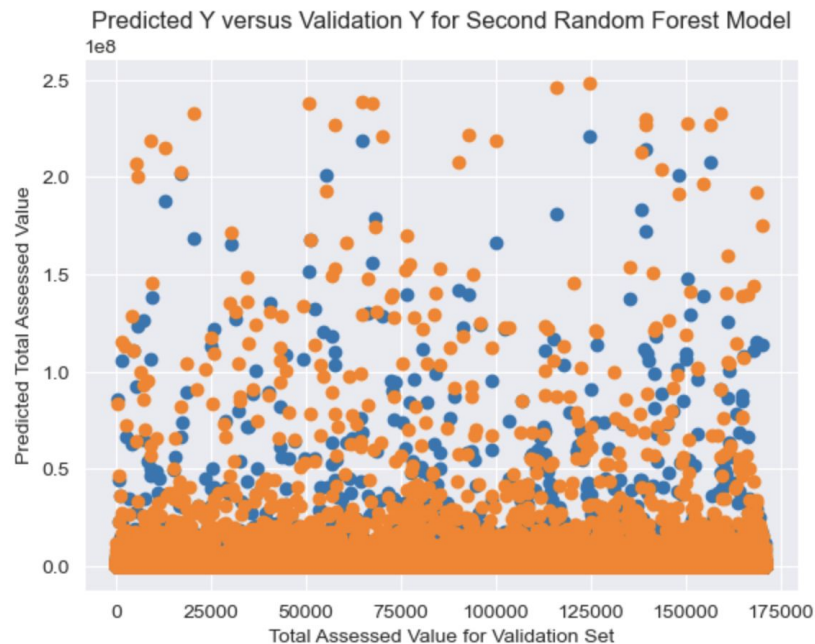
← Baseline

Random Forest Regressor

n_estimators = 10



n_estimators = 20




Conclusions

Random Forest Models are Unreasonably Effective



- We explored multiple flavors of both linear models and more complex NN models
- Our final Random Forest model does a reasonably good job of modeling our data
- It is vastly superior to the median baseline model and more basic linear models
- Our Random Forest Model with 20 n_estimators and default max_depth yields the best results among tested models

Loss of Explainability

- Due to the complexity of the relationships in our dataset we have had to sacrifice explainability
 - This makes our final model potentially useful for operations use as a predictor
 - But not useful as an explainable model to give us insight into what is happening in the SF housing market
- 


Known Biases

Aggregation



- Zip code level aggregation was necessary to provide a common scale for all of the available datasets
- We are losing some granularity and implementing implicit averaging by focusing on the zip code level

Missing Data

- In order to populate our zip code level records with missing data we have extrapolated neighborhood level data down to the zip code level
 - This further exacerbates the averaging the impact of the affected features
 - E.g. safety, neighborhood satisfaction, etc.
- 


Future Work

Better Data



- More granular data could yield better results
- Data not reliant on city appraisals
- More detailed data (currently neighborhood data), perhaps from custom surveys
- Incorporate additional data to better separate data points (zoning, unit level data, ownership status)

Better Models

- With more granular data we could attempt to find a fitting linear model
 - Alternatively, search for a more flexible FNN or CNN model to take advantage of more rich data
- 



Thanks!

Any Questions ?

Team Contributions



Notebooks:

Met up to collaborate on notebooks 4-5 times for work sessions.

EDA : Arsh, Anton

Data Cleaning: Priya, Sally, Jasmine

Data Merging: Priya, Sally, Jasmine

Model Making + Results: Priya, Sally, Jasmine

Cleaning up notebook : All



Slides:

Motivation: Arsh

Data (police, city survey, valuation):

- Introduce Raw data - Arsh
- Explain EDA -Arsh
- Cleaning - Sally
- Explain merging of data - Priya

Baseline Model (Linear Model 1,2 - Priya & Median - Jasmine)

Final Model (Random Forest 1,2 - Jasmine)

Conclusions: Anton

Contribution: ALL

