reonomy

Likely to Sell Prediction

Analytics in Practice Spring 2020

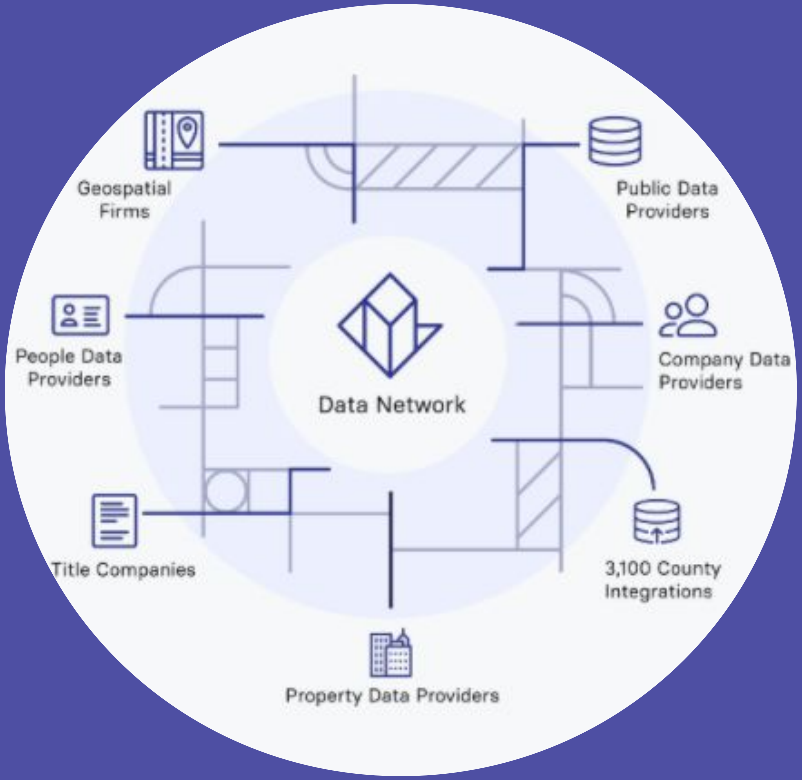
Aditi K. | Aneesh G.| Mihir A. | Metika S. | Yassine T.

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Unlocking Property Intelligence

- **Introduction**
- **Feature Engineering**
- **Model Exploration**
- **Results**

About



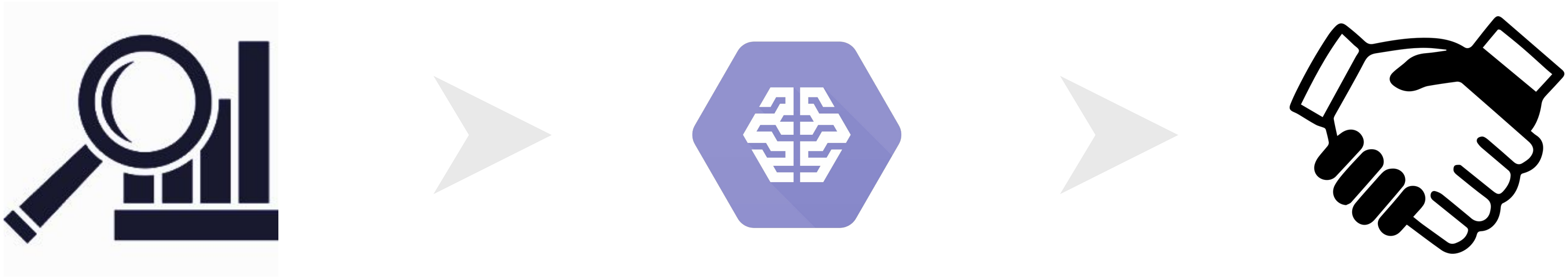
Reonomy **leverages big data**, partnerships and **machine learning** to connect the fragmented, disparate world of commercial real estate.

100 **sources of data**, including multiple public and proprietary data feeds and crowdsourced information, and then uses **artificial intelligence** to crunch it to provide market intelligence

The platform is used by developers, investors, acquirers and anyone else who works in the **area of commercial property**.

Problem Statement

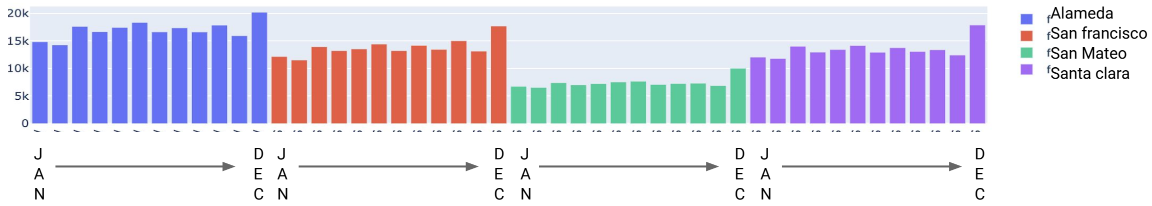
- **“Will that property be sold next year?” - Real Estate Associates**
- The **objective** of this project is to use **historical transactional** data to build a model that predicts if a property is l**ikely to be sold** next year
- Having this kind of **intelligence enables** the platform users (Realtors, brokers, etc) **act fast** and create value by approaching property owners well in advance.



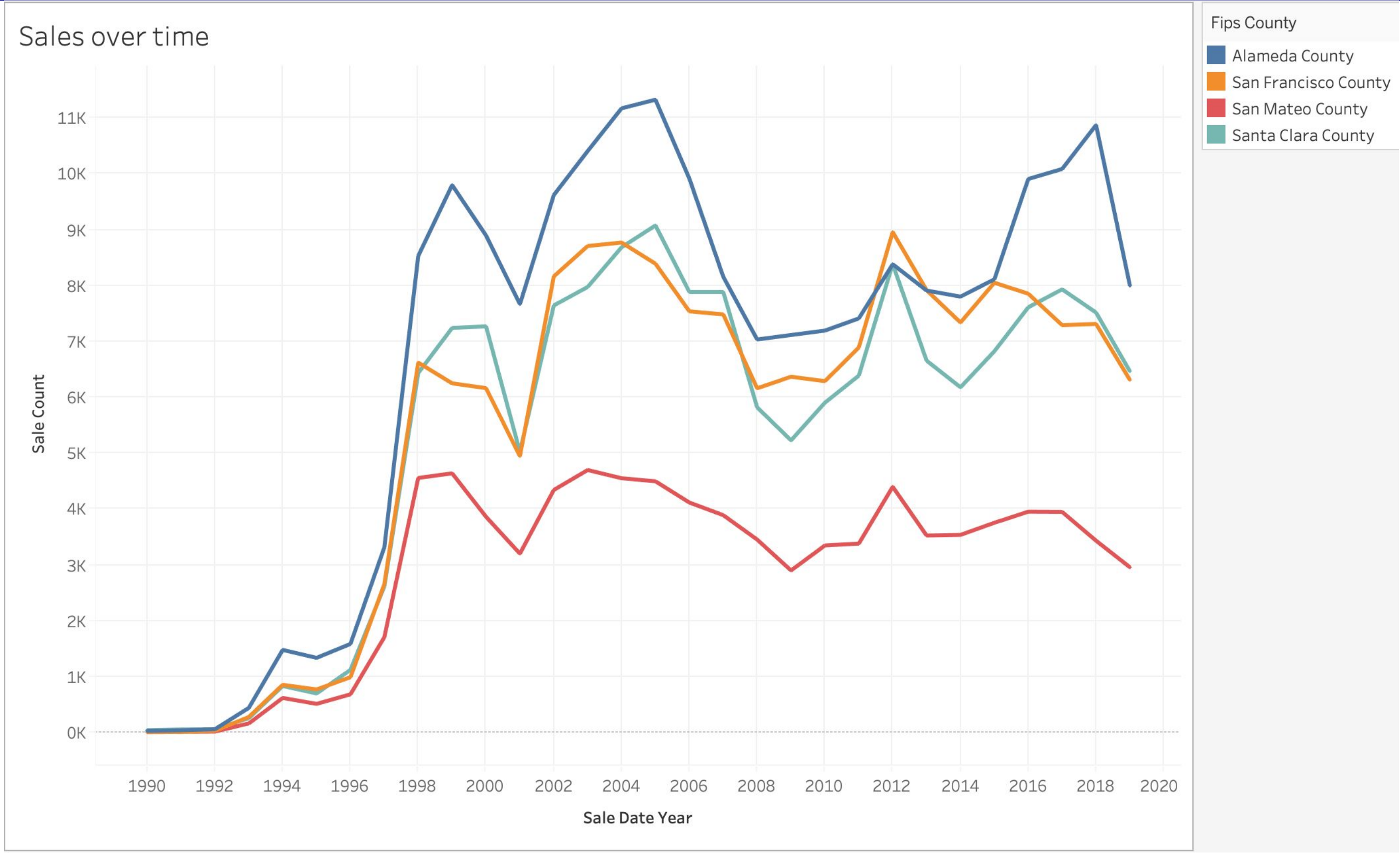
Dimensions to sale of a property

- The sale of the property depends on the **owner** as well as on a **buyer who is ready.**
- Factors that influence sales such as
  - **Age of owner**
  - **Marital Status**
  - **Mortgage period**
  - Cannot be incorporated in the model
- The property type comes with it's own seasonality behaviour.
- The data we have includes properties belonging to **Multi-family, Offices, Retail, Industrial** and **Hospitality class.**

Aggregate trend of Sales

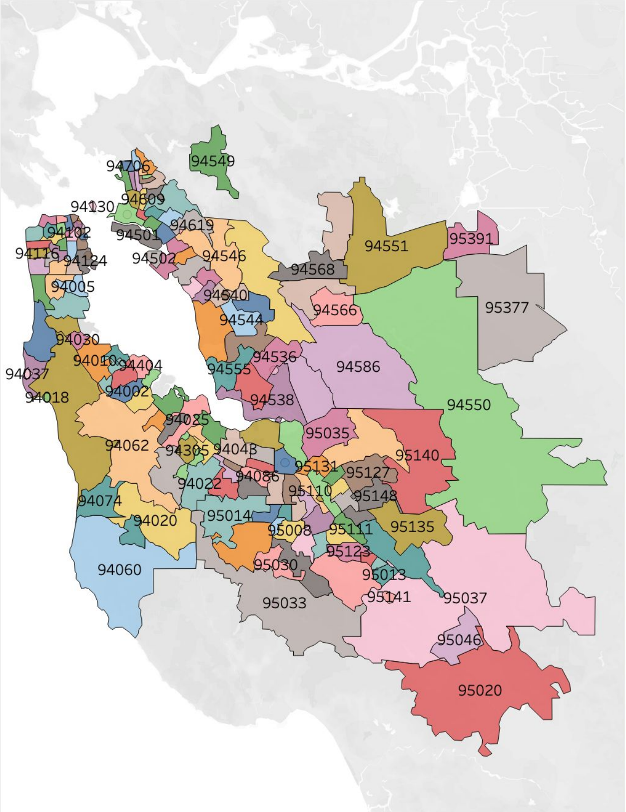


Granular look into seasonality

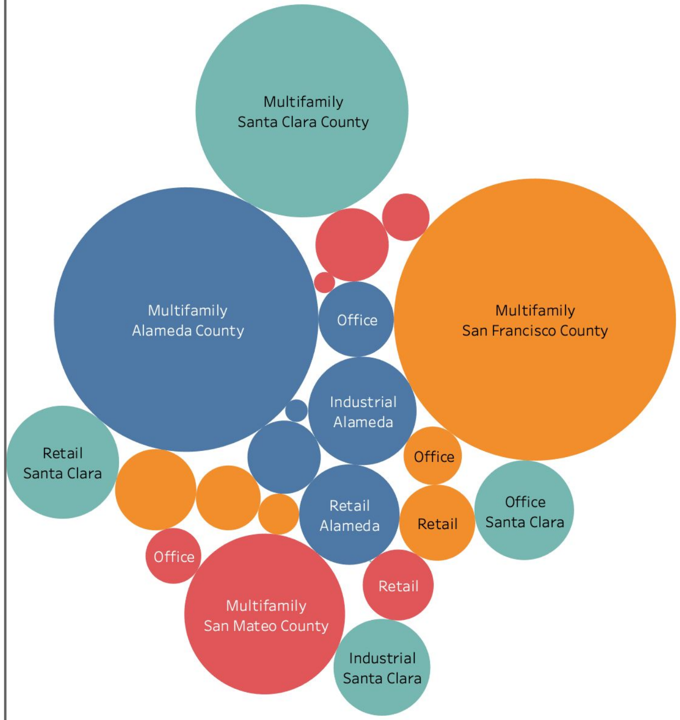


Geospatial Overview

ZipCode Level



CountyLevel



Unlocking Property Intelligence

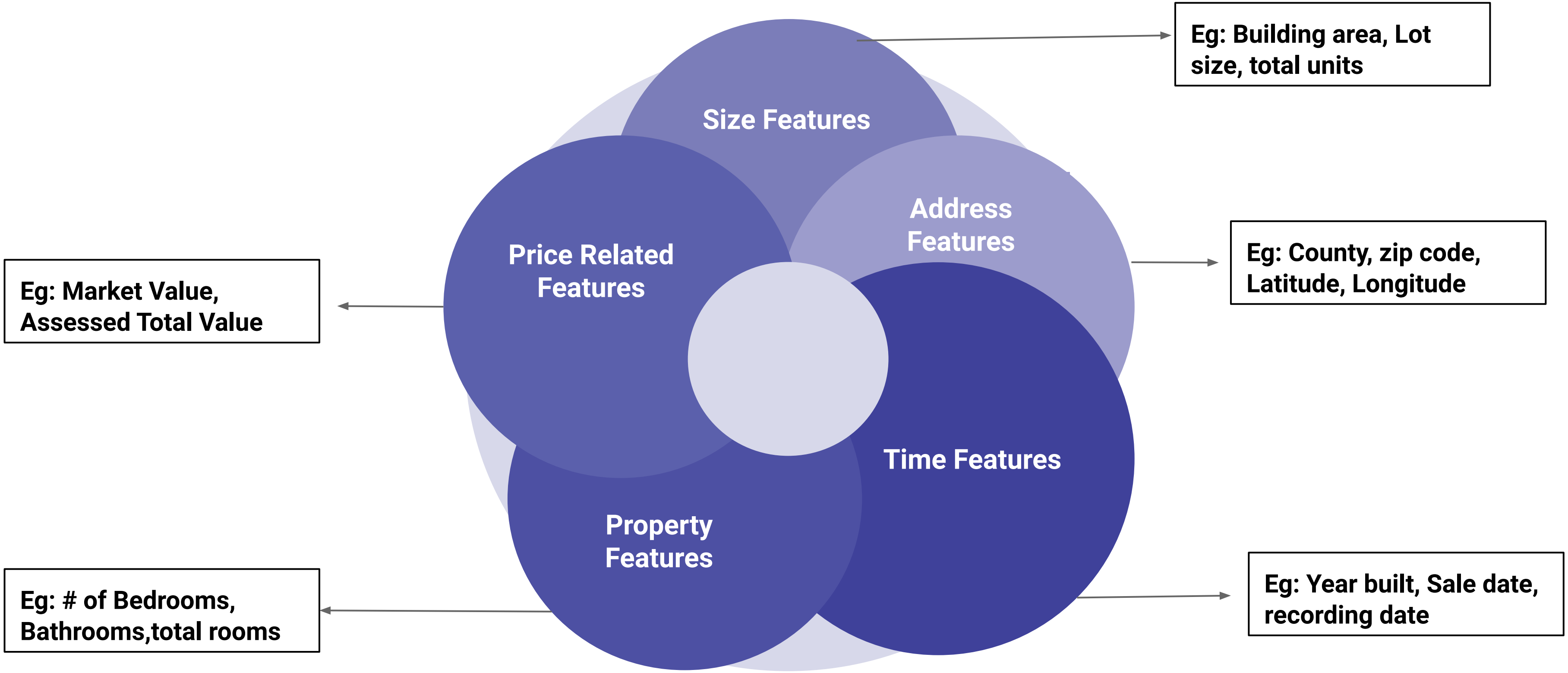
Introduction

Feature Engineering

Model Exploration

Results

Need for Feature Engineering



Need for Feature Engineering

Issue with *Static* data for our Problem Statement:

Features are constant throughout the sale history of the property

So we added some DYNAMIC Features!

Eg: Building area, Lot size, total units

Eg: County, zip code, Latitude, Longitude

Eg: Year built, Sale date, recording data

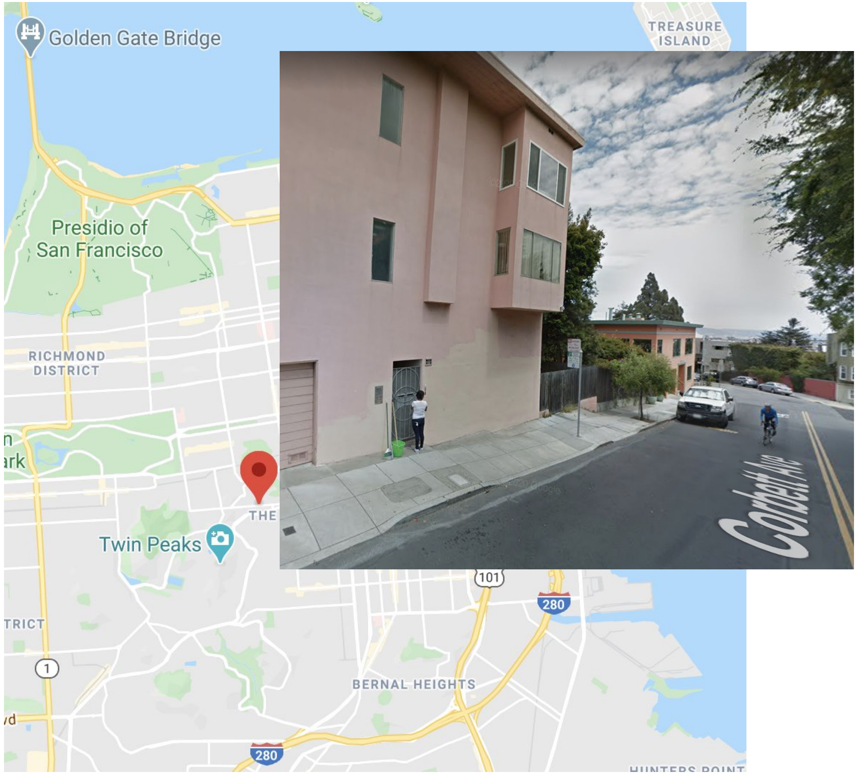
Eg: Market Value, Assessed Total Value

Eg: # of Bedrooms, Bathrooms,total rooms

Feature Engineering

Features (2018)

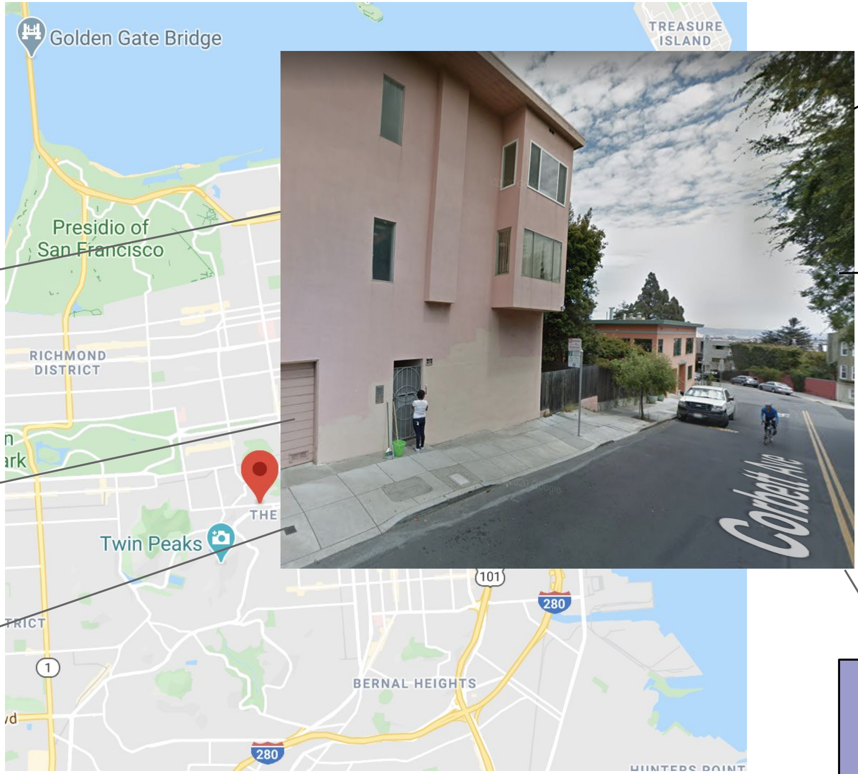
Multifamily property in SF county built in year 1961



Feature Engineering

Features (2018)

Multifamily property in SF county built in year 1961



Age of property in that Year

Ratio of sq ft area of property to avg. sqft are of sales in that Category, zip code, year

Ratio of Sales in Property Category, zip code in to tot. sale in that Year

# of total sales in that year for category, zip code

# of Years since Last Sale

Times Sold in Year | 1 yr before | • • • | 10 yrs before

Ratio of # of bedrooms to avg. # of bedrooms for sales in Category, zip code, Year

Data Map

Property _id	Property Features				Time features				Sold/Not Sold
	Average over				Number of times sold in				
	Market Value		Building Area		1 year	2 years	....	10 years	
	Category	County	Zipcode	City					
004dfdf-71	Multifamily	San Fransisco	94016	San Fransisco	0	1	....	3	0
004dfdf-82	Retail	Alameda	94501	Oakland	1	1	....	4	1
004dfdf-86	Retail	San Jose	94088	San Jose	0	0	....	2	1
004dfdf-87	Office	Santa Clara	94020	Palo Alto	0	1	....	3	1
004dfdf-88	Industrial	Santa Clara	94043	Sunnyvale	0	0	....	0	0
004dfdf-89	Mixed Use	Alameda	94501	Oakland	2	3	....	9	1
004dfdf-90	Industrial	San Fransisco	94016	San Fransisco	6	1	....	8	0

Unlocking Property Intelligence

Introduction

Feature Engineering

Model Exploration

Results

Model Selection

Gradient Boosting was chosen as the model of choice due to the following reasons:

- Logistic Regression showed that the relationship between the features and label was **non-linear**
- Tree based ensemble models were able to **capture the trend better** as they are **non parametric**
- Gradient Boosting takes into account the **unbalanced nature** of the data set
- The algorithm is sequential & provides with many weak learners which reach the optimal solution
- The **learning rate** of these learners can be **tweaked**
- It also gives a **feature importance graph** which can be used to derive business insights

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.
2. For $b = 1, 2, \ldots, B$, repeat:
   (a) Fit a tree $\hat{f}^b$ with d splits ($d+1$ terminal nodes) to the training data (X, r).
   (b) Update $\hat{f}$ by adding in a shrunken version of the new tree:
   $$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x).$$
   (c) Update the residuals,
   $$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i).$$
3. Output the boosted model,
$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x).$$

Model Fitting

Getting the data ready for modelling:

- The converted data set was highly imbalanced as properties were sold only a few times from 1990-2019
- The model was run on a subset of the data to get the top features which affect sale in the next year
- The top 10 features were used to create 10 year lags in order to make the data balanced and use the entire train data
- The train set from 2010-2017 was made balanced by downsampling the labels
- The validation set and test set were used to pick the optimal threshold

Train

2010-2017

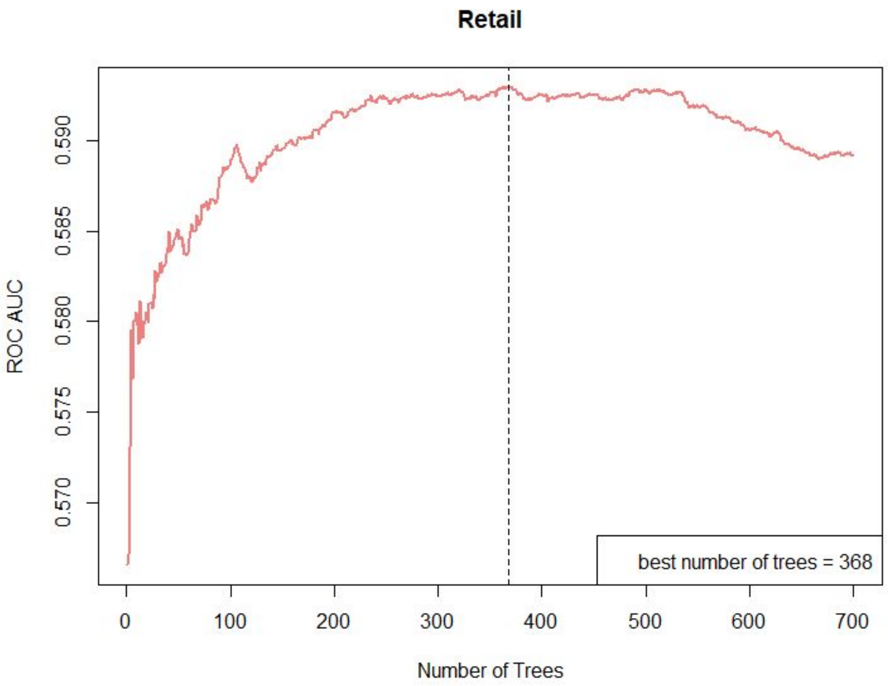
Validate

2018

Test

2019

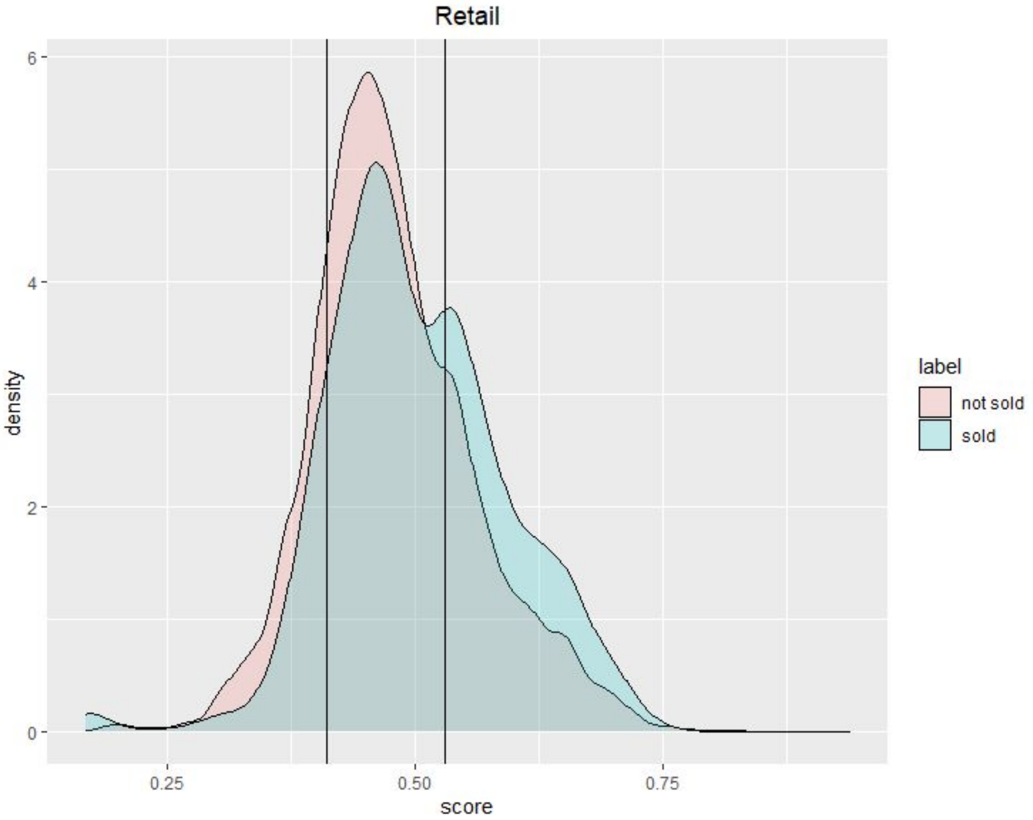
Parameter Tuning



To avoid **overfitting** we choose the **number of trees** that maximizes AUC on the validation set.

For **Retail**, this is **368** Trees.

Best Threshold



Trade-off between **precision** and **recall**

higher threshold:
Higher confidence
Lower coverage

lower threshold:
Lower confidence
Higher coverage

Unlocking Property Intelligence

Introduction

Feature Engineering

Model Exploration

Results

Results

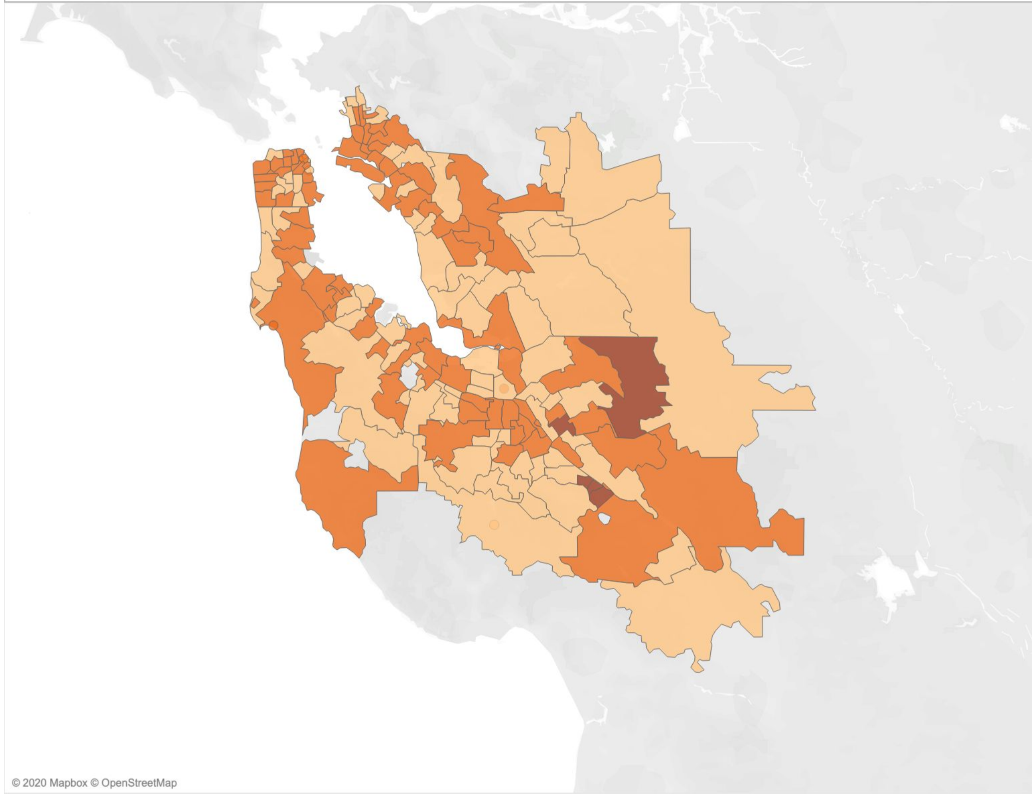
Type	Trees	AUC	Threshold	Recall	Precision	Important Features
Multifamily	997	0.57	**0.48** | 0.62	**46%** | 11%	**12%** | 20%	**years since last sale** baths/avg_baths_year_category_zipcode category_zipcode_sales/total_sales
Retail	368	0.59	**0.55** | 0.64	**25%** | 9%	**14%** | 21%	**years since last sale** sqft/avg_sqft_year_category_zip category_zipcodes_sales/total_sales
Industrial	326	0.61	**0.52** | 0.61	**36%** | 10%	**12%** | 19%	**years since last sale** category_sales/properties_category sqft/avg_sqft_year_cat_zip
Office	306	0.60	**0.54** | 0.61	**28%** | 6%	**18%** | 30%	**years since last sale** category_zipcode_sales/total_sales sqft/avg_sqft_year_category_zipcode
Mixed Use	44	0.55	**0.49** | 0.54	**34%** | 12%	**13%** | 16%	**years since last sale** sqft/avg_sqft_year_category_zipcode rooms/avg_rooms_year_category_zipcode
Hospitality	69	0.61	**0.49** | 0.59	**33%** | 4%	**14%** | 19%	**years since last sale** baths/avg_baths_year_category_zipcode rooms/avg_rooms_year_category_zipcode

Results

Type	Trees	AUC	Threshold	Recall	Precision	Important Features
Multifamily	997	0.57	0.48 | **0.62**	46% | **11%**	12% | **20%**	**years since last sale** baths/avg_baths_year_category_zipcode category_zipcode_sales/total_sales
Retail	368	0.59	0.55 | **0.64**	25% | **9%**	14% | **21%**	**years since last sale** sqft/avg_sqft_year_category_zip category_zipcodes_sales/total_sales
Industrial	326	0.61	0.52 | **0.61**	36% | **10%**	12% | **19%**	**years since last sale** category_sales/properties_category sqft/avg_sqft_year_cat_zip
Office	306	0.60	0.54 | **0.61**	28% | **6%**	18% | **30%**	**years since last sale** category_zipcode_sales/total_sales sqft/avg_sqft_year_category_zipcode
Mixed Use	44	0.55	0.49 | **0.54**	34% | **12%**	13% | **16%**	**years since last sale** sqft/avg_sqft_year_category_zipcode rooms/avg_rooms_year_category_zipcode
Hospitality	69	0.61	0.49 | **0.59**	33% | **4%**	14% | **19%**	**years since last sale** baths/avg_baths_year_category_zipcode rooms/avg_rooms_year_category_zipcode

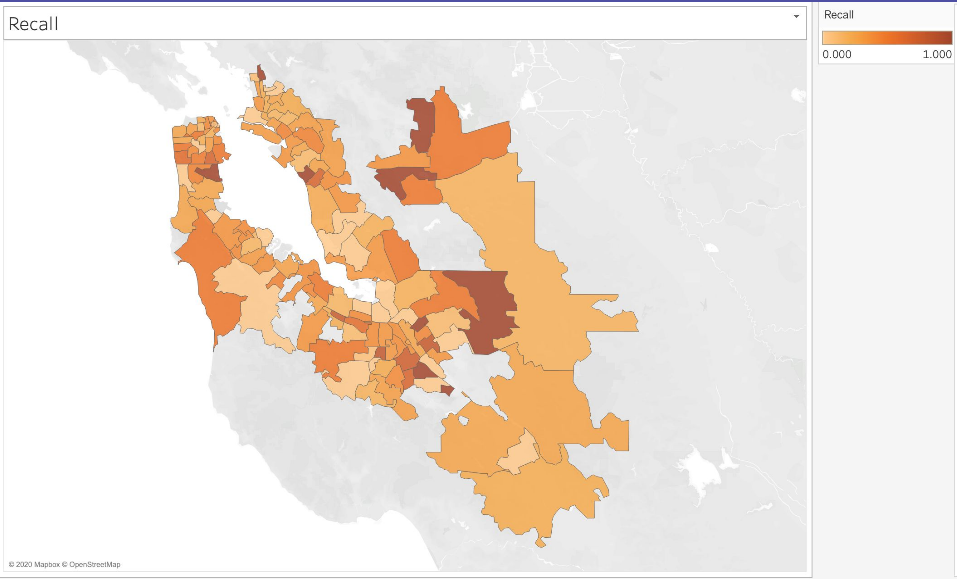
2019 Prediction

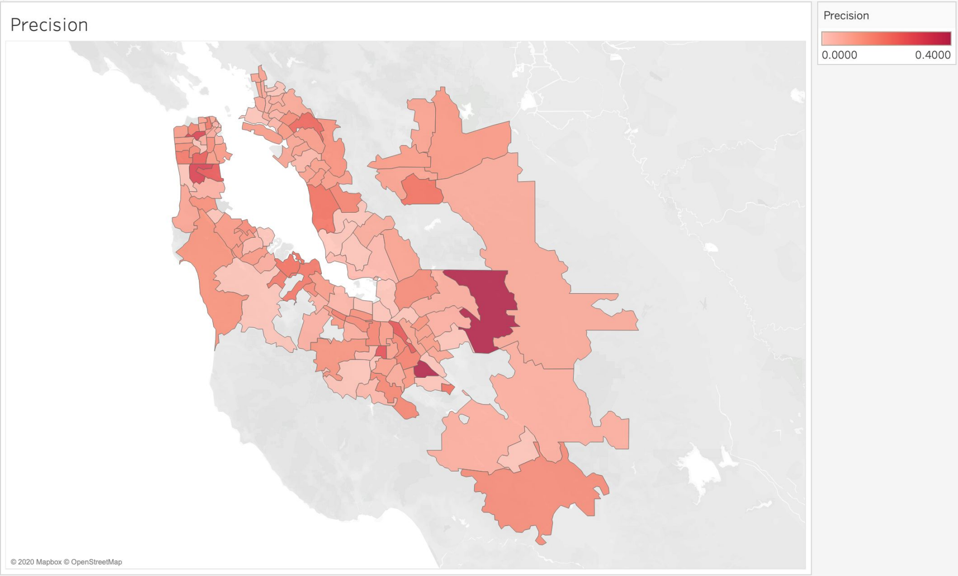
Confidence of Sale



- Highly likely to sell
- Likely to sell
- Not likely to sell

2019 Prediction





- Higher recall & precision in densely populated areas
- zip codes can be used as a targeting strategy for brokers

Thank you

Team REO3