The Investment Property Deal You Will Never Miss

Mark Zhao, 02/14/2020

Executive Summary

Using scraped and cleaned data from Redfin and Airbnb, I designed a machine learning model that can predict home price and rental income.

Given a home’s features like beds, baths, type, and location, the model is able to **predict the Return on Investment (ROI)** and **find the best deals** of investment property.

Key Findings

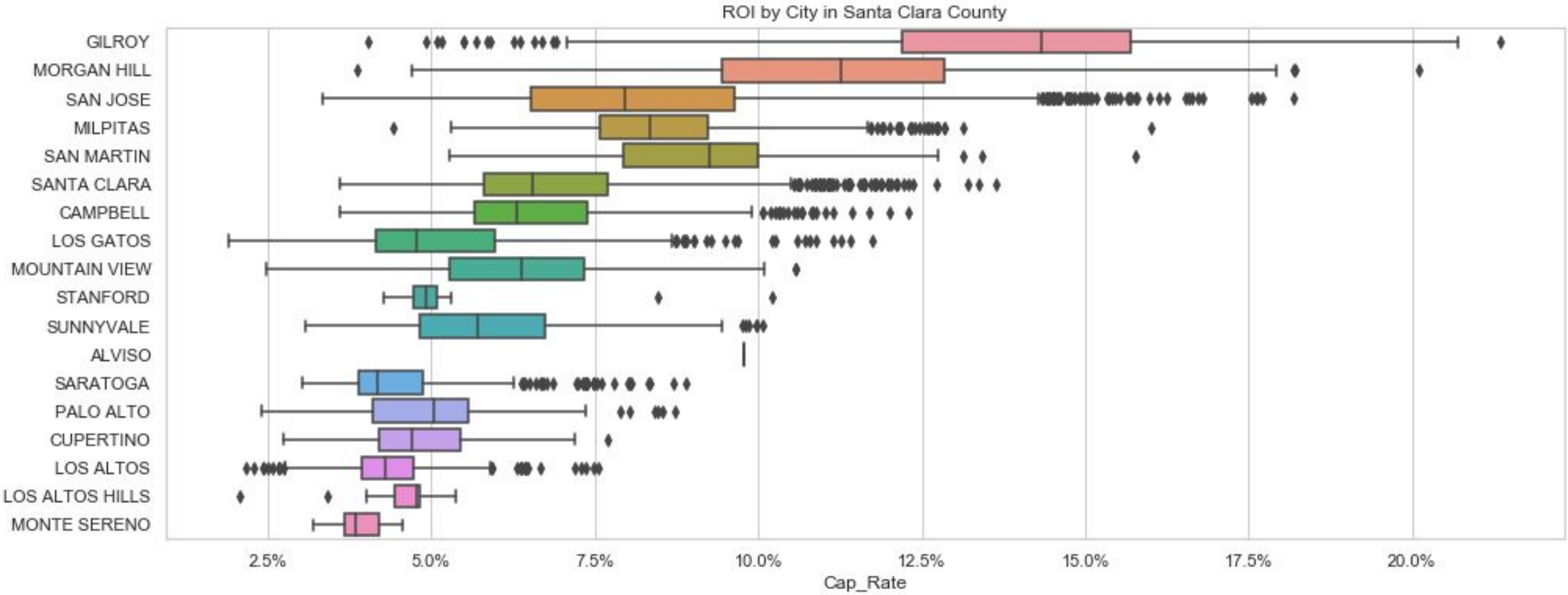
- Data suggest that as an rental property investor, you should avoid neighborhoods with good schools and focus on the ones with great location and size.

- Linear regression, LASSO, and Random Forest models were trained and tested with Random Forest being the winner yielding test R^2 = 92.5% for home price and 75.9% for rental income.

- Compared to the median ROI which is around 7.5%, the model can identify investment properties with **yields > 15%, or twice the return of the median**.

- The 90th percentile of all predicted cap rates is 11.6%, about 4% higher than the median. Compounded for 10 years with an initial investment of $1 million, **this means $900,000 more wealth** which makes quite a big difference.

Return on Investment by City



Problem Statement

Many affluent families are considering buying investment properties for rental income and wealth building.

You wonder whether to buy a 5 bedroom single family house in a nice neighborhood with good school districts or a 3 bedroom townhouse close to downtown where big companies located at.

There’s no obvious way to make a decision without data. You get different information from different people and have no clues which one to choose.

Solution

Introducing a revolutionary tool that could change the game of rental investment. Imagine you can browse thousands of homes that are sorted by return on investment and tell you which are the best deals to grab?

This is a whole new experience that gives you many more choices, more flexibility, and more information so you can get a clear and personalized recommendation of investment property.

Data Collection

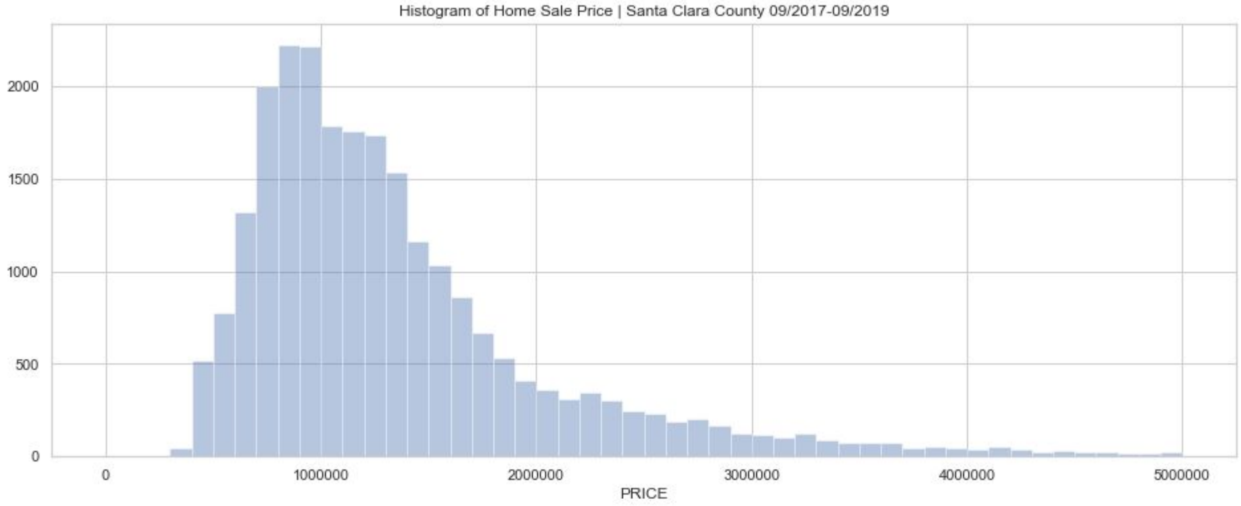
- Home Price and Features: scraped from Redfin for Santa Clara County (SCC) in San Francisco Bay Area. There are round 30,000 observations sold from 09/2017 to 09/2019 with home price as target and features like beds, baths, sqft, lot, year, property type, HOA, sale date, and coordinates.
- Airbnb Rental Price and Features: scraped data with around 7000 listings as of 07/09/2019 in SCC. There are more than 100 features including home characteristics, reviews, and amenities.
- Landmark Coordinates: 17 landmarks' coordinates data were collected from Google Map to calculate the geodesic distance from each property/listing to these landmarks including the big tech companies, downtowns, and Levi's stadium etc.
- School Scores: percentage standard met and above for Math and English tests were extracted from CAASPP and were averaged on the zip code level.
- Crime: The violent crime (including murder, rape, robbery, and aggravated assault) and property crime (theft, burglary, and arson) index data were collected from bestplaces.net on zip code level.

Data Cleaning

- Multi-Family (5+ Units) and land were removed because they tend to be a commercial office or big apartments with extremely high prices. Only property types SFH, Townhouse, Condo and Multi-Family (2-4 Unit) were included.
- To test the hypothesis that home sale prices are affected by seasons, 4 dummy variables extracted from the sale date were added to indicate whether the sale happens in Q1, Q2, Q3, or Q4.
- Observations with price < 100K, or beds/baths < 1, or square feet < 10 were removed due to data errors.
- Since beds, baths, square feet, lot size, year built are the main features, observations with missing values were removed (<5% of total population).
- For HOA/MONTH, 70% of the observations are NaNs and the median HOA by property type were filled.

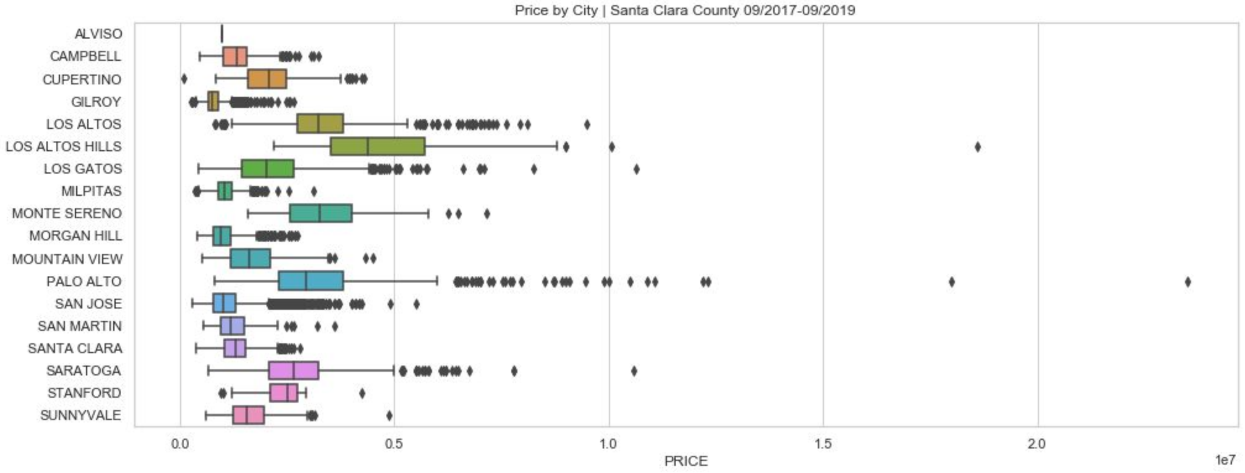
Exploratory Data Analysis

The distribution of home sale price is normal but skewed to the right, suggesting outliers with extremely high price.



Home Price by City

Neighborhoods like Los Altos, Los Altos Hills, Palo Alto, Saratoga, and Monte Sereno exhibit higher median price and outliers which is consistent with the expectation

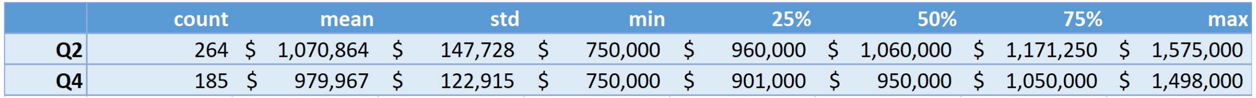


Do You Get a Deal in Q4?

Anecdotes suggest if you buy a home at the end of the year you tend to get a deal. On the other hand, you pay a premium (probably through a bidding war) in late spring to early summer for your home. Is it a myth or is it really true?

Null Hypothesis: There are no differences in price between homes sold in Q2 and Q4.

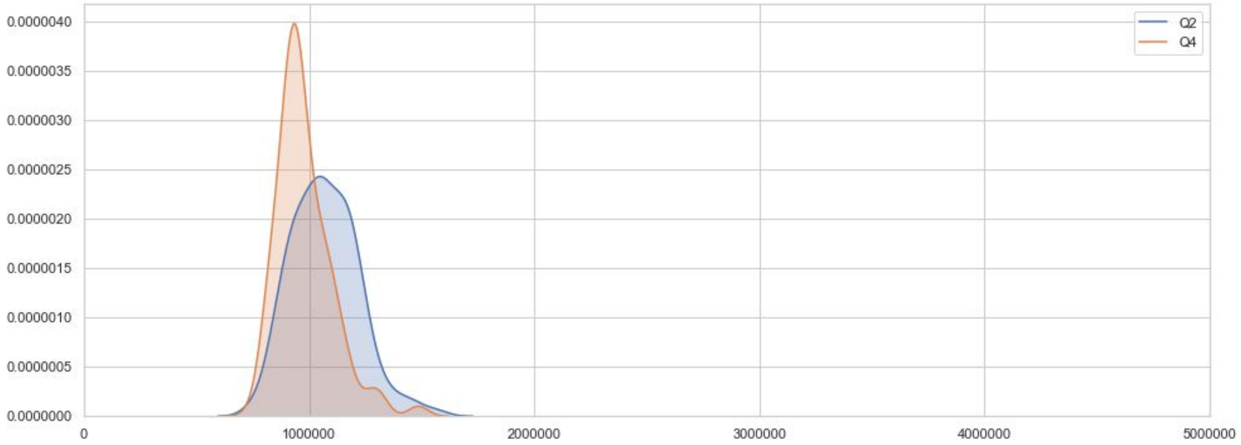
If we look at price by quarters, the median, 25th and 75th percentile price is higher for Q2 than any other quarter (see Figure 5). This suggests the anecdotes might be true.



	count	mean	std	min	25%	50%	75%	max
Q2	264	$ 1,070,864	$ 147,728	$ 750,000	$ 960,000	$ 1,060,000	$ 1,171,250	$ 1,575,000
Q4	185	$ 979,967	$ 122,915	$ 750,000	$ 901,000	$ 950,000	$ 1,050,000	$ 1,498,000

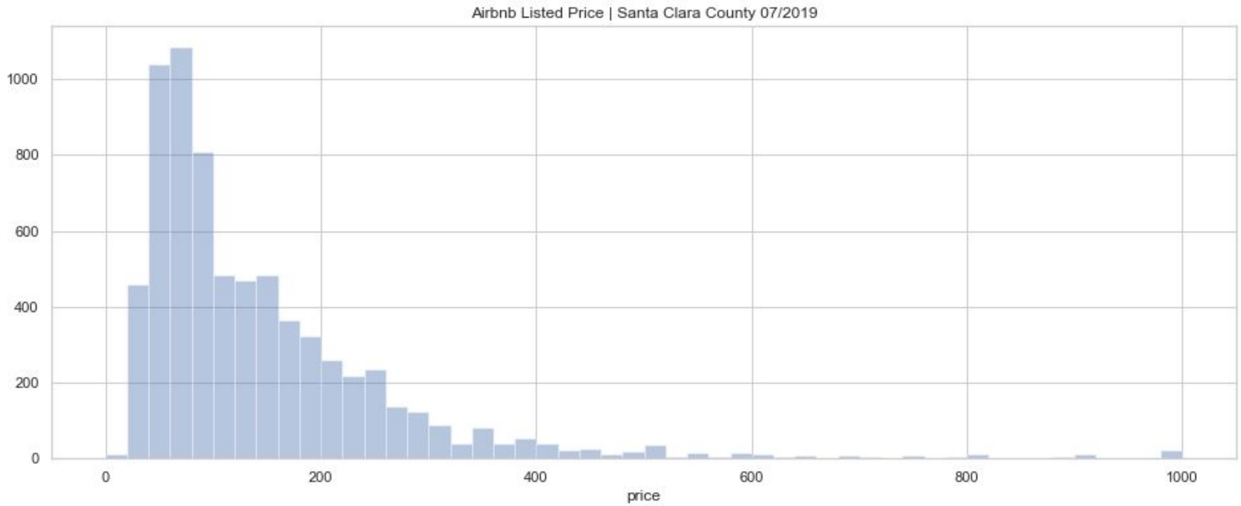
Q4 Seems to be a Bargain Season

The probability density functions of the two samples also suggest Q4 is cheaper. The t value is 7.09 and the p-value is close to 0. The null hypothesis is rejected with a level of significance level of 0.01.



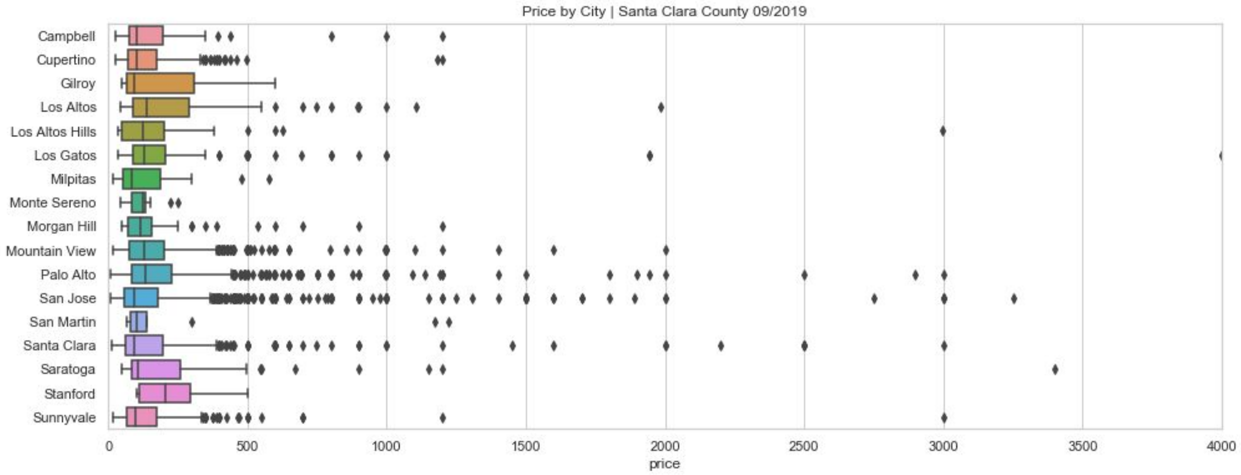
A $10,000/night Airbnb Room

The distribution of home sale price is normal but skewed to the right, suggesting outliers with extremely high price. The max one-day Airbnb price is $10,000! (the figure below was cutoff at $1000)



Location, Location, Location

Mountain View is not the most expensive neighborhood in terms of home price but charges one of the highest Airbnb price. This might suggest that location is very important for Airbnb prices.



Machine Learning

The goal of the in-depth analysis is, given the features of a potential property (such as number of beds/baths, coordinates, etc.) to predict

1) **how much to buy the property**,

2) **how much rental income to get** if listed on Airbnb,

And based on 1) and 2) to calculate the **capitalization rate** (ROI without considering mortgages).

Home Price Train Test Split

Home prices with more than $3 million tend to be outliers and were therefore removed from the analysis.

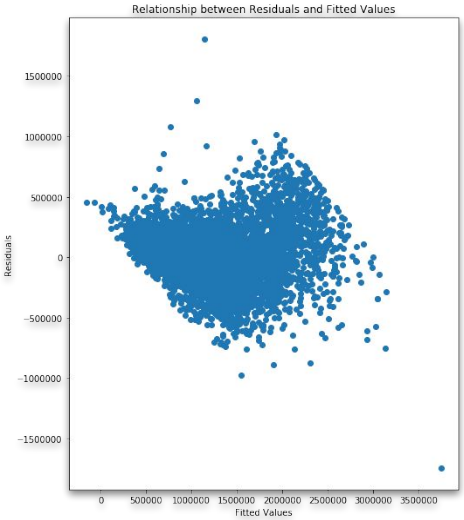
There are 16141 observations (70%) in the training data and 6918 (30%) in the testing data with 110 features.

List of features: Beds, Baths, Square feet, Lot size, Year, HOA, Latitude, Longitude, Dummy variables for 1) the quarter of the sale 2) city 3) zip code 4) property type, Distances to different locations such as (Google, Apple, San Jose Downtown, Palo Alto Downtown, etc.), School, Crime

Home Price - Linear Regression and LASSO

A simple linear regression model shows a quite decent performance with around 84% R^2. To prevent overfitting and reduce features, a LASSO model was tested with 8 variables eliminated and a final R^2 of 83%. Square feet turns out to be the most important feature which is not a surprise.

However, the relationship between residuals and fitted values shows the heteroscedasticity of the errors.



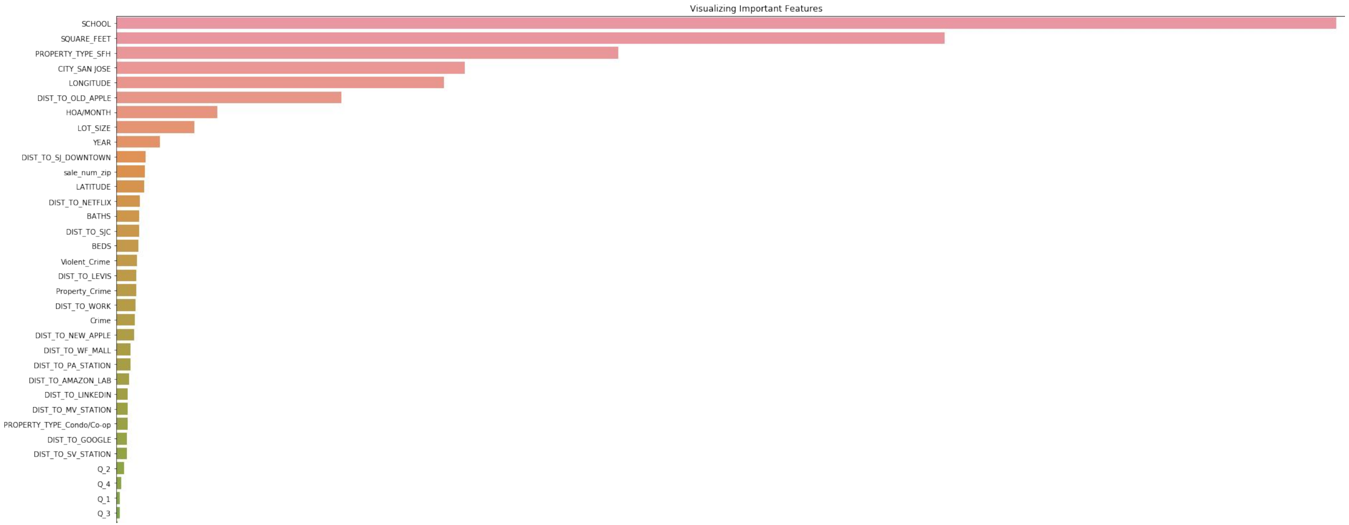
Home Price - Random Forest

A random forest model with n_estimators=100 and max_depth=20 was tested and the testing score improved to around 92% R^2. The RMSE is around 148000 and the median absolute error is around $70,000.

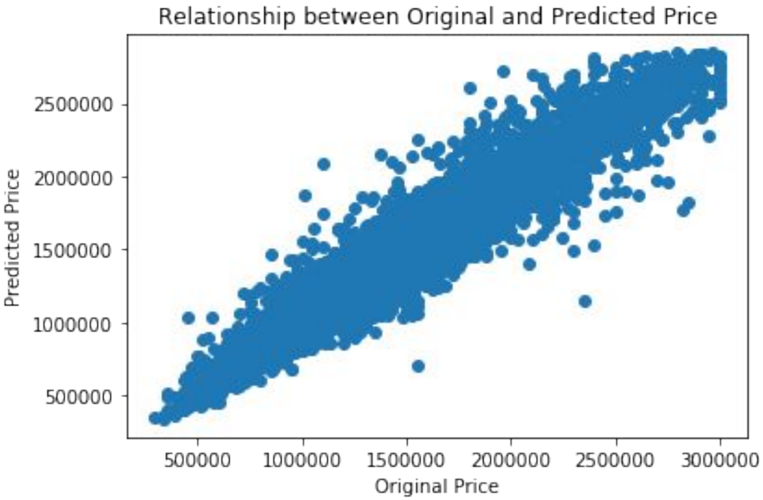
Features ordered by importance with the most important ones include **School, Square feet, Property type SFH and then some location features**. This is in line with my hypothesis that schools and the size of the house are important when purchasing homes.

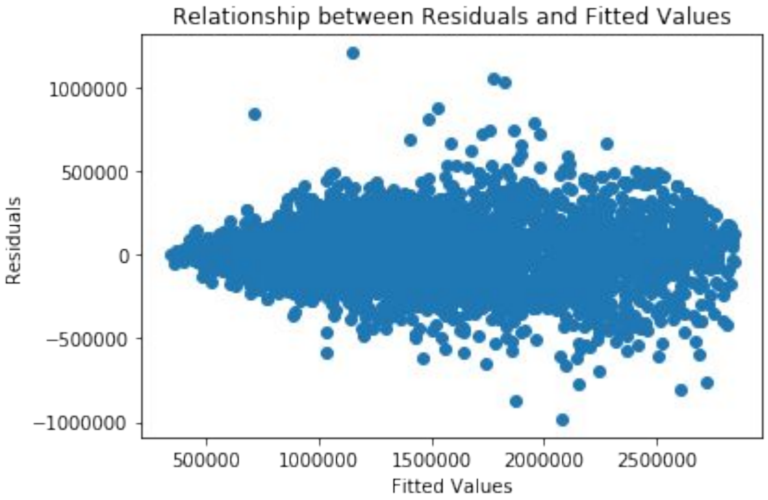
The errors are close to randomly distributed and there are very few outliers.

School and Size Matter



92% R^2! 8% is from Home Pictures?


Relationship between Original and Predicted Price


Relationship between Residuals and Fitted Values

Rental Income Train Test Split

Training data and testing data were split from the original data with 70% and 30% respectively. Given the list of features in the Airbnb data that overlapped with the Redfin data: Beds, Baths, Latitude, Longitude, Dummy variables for 1) city 2) zip code 3) property type, Distances to different locations, School, and Crime, supervised learning models were tested.

One more dummy variable is added which indicates the entire home is listed. This can be personalized further if the investor wants to list part of the home for rent.

Target variable is the average listing price times the number of booked days in the next 30 days. The goal is to find the "equilibrium" rental price, and so listings with extreme prices and/or with very few bookings were removed with filters of vacant days in the next 30 days <=7, the number of reviews in the last 12 months >=5, and the average daily price <=$500. In the end, there are 1655 observations in the training data and 710 observations in the testing data.

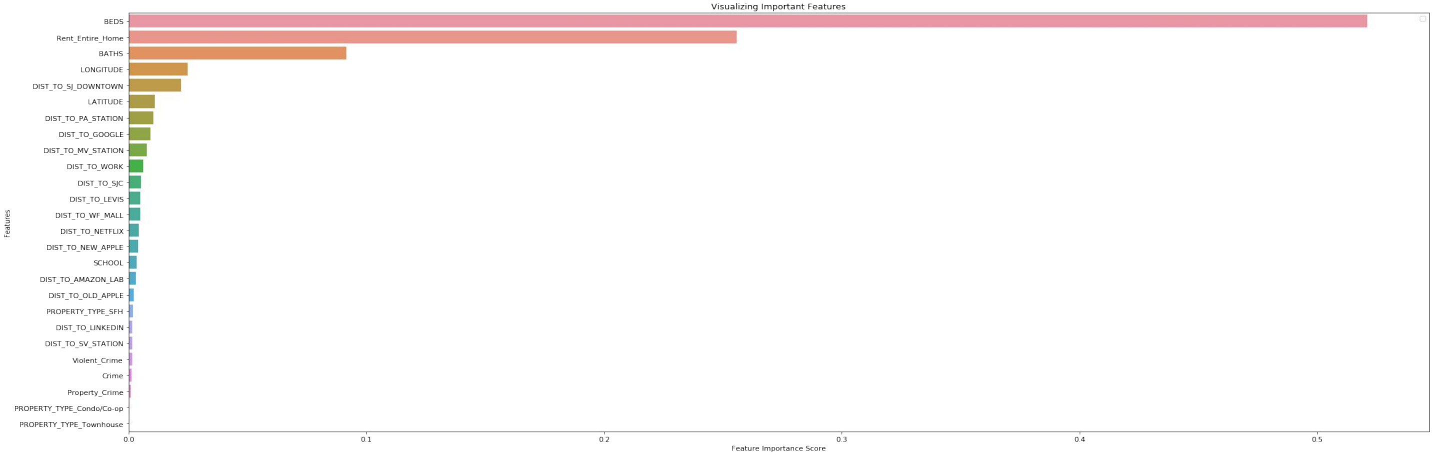
Rental Income - Random Forest

A random forest model is tested with a testing score around 76% R^2 and the median absolute error $565.

The most important features include **beds, rent_entire_home, and baths** which seem reasonable. Interestingly, school and crime are less important than location features. Renters mostly seek shorter-term residence and care less about the school and crime and more about location.

Cross-validation didn’t improve the performance much.

Airbnb Renters don’t Care School and Crime



Return on Investment - The Holy Grail

The capitalization rate is monthly income * 12/ property purchase price without considering taxes, mortgages, expenses, home price appreciation for simplicity.

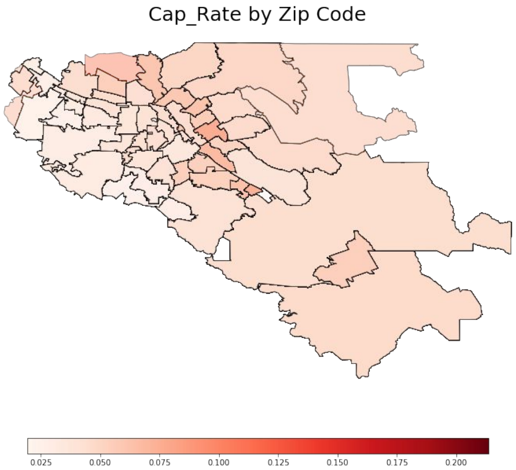
The average predicted cap rate is 7.5% which is quite reasonable. **The 90th percentile of cap rates is 11.6%, about 4% higher than the median**. Compounded for 10 years with an initial investment of $1 million, this means **$900,000 more wealth** which makes a big difference.

A better deal (high ROI) can be found in neighborhoods with below-average school but in a great location (close to work hubs) and the initial investment is lower (cheaper to buy).

Buy Low, Rent High

The deeper the red, the higher the cap rate and the better the deal. Buying homes in nicer neighborhoods such as Cupertino, Los Altos and Saratoga might give you a higher dollar amount of monthly income but you also paid a (much) higher price for the property and thus they're the worst deals in terms of lower cap rate.

Another interesting observation is Palo Alto: although the home price is quite high there it's also close to work and school so the cap rate is one of the highest in the nicer neighborhoods.



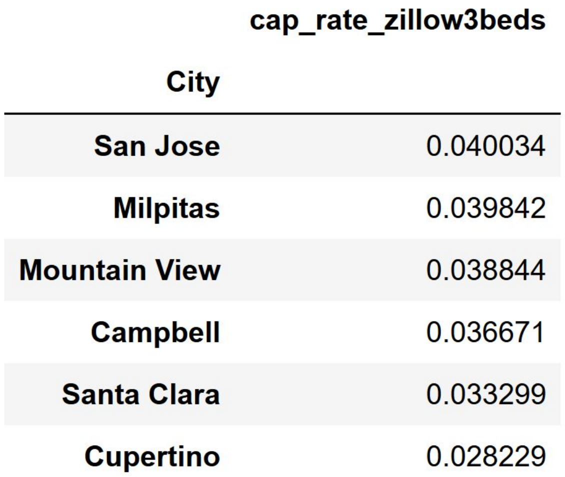
Source: Redfin, sccgov

What about the Long-term Rental?

So far I’ve been using Airbnb rental price to calculate the rental income. Although Airbnb is getting more and more popular, what about the traditional long-term rental?

To test this, a different data source is analyzed: Zillow Rental Index which has average rental income. The data is on the zip code level and the income is average income so it’s not as personalized as using the Redfin data but nonetheless still good data to confirm my story.

The results suggest a very similar story to the Airbnb data which is comforting.



	cap_rate_zillow3beds
City	
San Jose	0.040034
Milpitas	0.039842
Mountain View	0.038844
Campbell	0.036671
Santa Clara	0.033299
Cupertino	0.028229

Caveat

The number of observations for Airbnb data is relatively small. Data from other counties and cities can be merged to get more observations.

To get a more realistic calculation of the capitalization rate, monthly operating costs can be estimated and added in the analysis. Plus, the flexibility of purchasing properties with loans can be added as well.

There are apparently some important missing features such as the home's pictures, home amenities, and finer school and crime data. For example, two homes can look exactly the same based on the current features but one is newly renovated and the other is not. The quality of the house can play an important role in both purchase price and rental price.

Conclusion

Although there are still rooms to improve, this model is a good starting point for people to invest in rental property. Give some basic features of a home, the model is able to predict the purchase price and rental income.

Based on the calculated capitalization rate together with the monthly cash flows and initial investment amount, people can make a much more informed decision in purchasing an investment property.

Hope this can help you purchase an investment property. It certainly helped me.